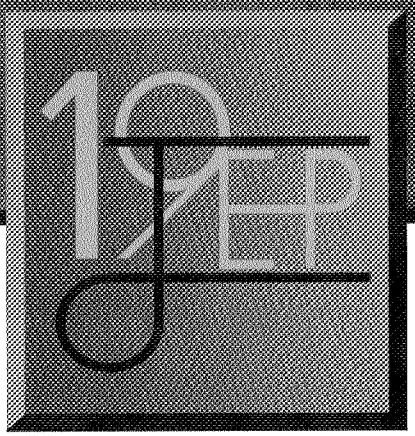
UNIVERSITE LIBRE DE BRUXELLES

INSTITUT DE PHONETIQUE



ACTES

19èmes JOURNEES D'ETUDE SUR LA PAROLE

Bruxelles, du 19 au 22 mai 1992

UNIVERSITE LIBRE DE BRUXELLES
INSTITUT DE PHONETIQUE
LABORATOIRE DE PHONETIQUE EXPERIMENTALE

19es JOURNEES D'ETUDE SUR LA PAROLE

Organisées sous les auspices
de la Société Française d'Acoustique
et de l'Association Belge des Acousticiens

Université Libre de Bruxelles
Institut de Phonétique (CP 110)
avenue F.D. Roosevelt 50
B-1050 Bruxelles
Belgique
Tél. : +32 2 650.20.08
+32 2 650.20.10
Fax : +32 2 650.20.07

n° dépôt légal : 1992-6332-1

COMITE SCIENTIFIQUE

PRESIDENTS

Henri Méloni, Université d'Avignon, Président du Groupe de la Communication Parlée
Maximilien Wajskop, Université Libre de Bruxelles

MEMBRES

Christophe d'Alessandro
Régine André-Obrecht
Marc Baudry
Christian Benoît
Jean-François Bonnot
Louis-Jean Boë
Jean Caëlen
Gérard Chollet
Jean-Marie Hombert
Paul Jospa
Jean-Paul Lefevre
Laurent Miclet
José Morais
Jean Schoentgen
Bernard Teston
Jacqueline Vaissière
Nadine Vigouroux

COMITE ORGANISATEUR

Raoul De Guchteneere
Paul Jospa
José Morais
Michel Piterman
Jean Schoentgen
Max Wajskop

avec la collaboration de Nathalie Bronkart et Joëlle Herzet

Université Libre de Bruxelles - Institut de Phonétique (CP 110) - avenue F.D. Roosevelt 50 - B-1050 Bruxelles - Belgique - Tél. :+32 2 650.20.08 - Fax :+32 2 650.20.07

Nous remercions :

- l'Université Libre de Bruxelles,
- le Commissariat général aux Relations internationales de la Communauté française de Belgique,
- le Ministère de l'Education, de la Recherche et de la Formation de la Communauté française de Belgique,
- et la Commission communautaire française.

pour l'aide qu'ils nous ont accordée et sans laquelle ces Journées n'auraient pu être organisées.

T A B L E D E S M A T I E R E S

2 - Phonétique
Président de séance : *M. ROSSI*

3 - Synthèse et conversion texte - parole
Président de séance : *R. CARRE*

Conférence plénière

4 - Perception I
Président de séance : *F. LONCHAMP*

INTRODUCTION

Depuis plus de 20 ans nos Journées d'Etude sur la Parole réunissent, dans une ambiance à la fois chaleureuse et laborieuse, de nombreux spécialistes qui viennent présenter et confronter en français leurs travaux concernant les multiples aspects de la communication orale (production, perception, apprentissage, représentation, pathologie, reconnaissance, synthèse, etc.). Cette année encore, le nombre exceptionnel des articles proposés, l'intérêt et l'étendue des thèmes abordés, la qualité des participants ainsi que le soin apporté à l'organisation scientifique et matérielle contribueront au succès et à la fécondité de ces 19èmes rencontres.

Les nombreux chercheurs qui participent régulièrement à ce congrès francophone relèvent de disciplines très diverses. La parole constitue pour certains le champ privilégié de leurs travaux et pour d'autres, un domaine particulier d'applications visant à mettre en valeur ou à expérimenter des méthodologies et des techniques développées dans un cadre plus général. Cette apparente hétérogénéité des motivations représente une richesse qu'il convient de préserver et d'encourager car elle garantit le dynamisme, la variété, l'étendue et la profondeur de nos recherches.

Pour chacune des disciplines concernées par l'étude de la parole, des sollicitations multiples incitent les chercheurs à publier dans de nombreux et prestigieux congrès et revues internationaux spécialisés. Les J.E.P. demeurent cependant un lieu de rencontres scientifiques très prisé, où la langue et la spécificité du domaine permettent des échanges approfondis sur l'ensemble des thèmes de la communication parlée. En outre, la présence simultanée de nombreux chercheurs débutants et de chercheurs depuis longtemps reconnus, confère à cette manifestation de qualité un indispensable caractère pédagogique.

Depuis sa création, le Groupe de la Communication Parlée a résolument - bien que sans exclusive - opté pour la francophonie. Aujourd'hui, des enjeux culturels, sociaux et économiques fondamentaux nous incitent à affermir cette position afin que le français demeure une langue vivante, parlée et comprise par le plus grand nombre d'individus et de machines. La coïncidence des 19èmes J.E.P. avec le 30ème anniversaire de la création de l'Institut de Phonétique de Bruxelles nous fournit l'occasion de réaffirmer notre attachement à cette orientation.

Au printemps 1994, nous célèbrerons avec éclat le 20ème anniversaire des J.E.P. en plein essor dont on peut attendre les plus grandes félicités. L'enthousiasme des nombreux prétendants à l'organisation de cette fête laisse augurer de sa pleine réussite.

Henri MELONI

PREFACE

1973, 1984, 1992. Pour la troisième fois, l'Institut de Phonétique de l'Université Libre de Bruxelles accueille les Journées d'Etude du Groupe de la Communication Parlée. Les 19es JEP coïncident avec le 30e anniversaire de l'Institut, fondé durant l'année académique 1962-63, comme les 4es et les 13es JEP coïncidèrent avec les 10e et 20e anniversaires de l'Institut. Ces rapprochements symboliques indiquent que la vie et l'histoire de notre Laboratoire de Phonétique expérimentale furent souvent en symbiose avec le développement du Groupe de la Communication Parlée.

Le Groupe de la Communication Parlée est désormais une institution adulte qui a joué un rôle de premier plan non seulement dans le développement des recherches françaises mais aussi dans l'élaboration du maillage européen des laboratoires, le renforcement des relations internationales au sein de la communauté parole, la création de *Speech Communication* et la fondation d'ESCA. La surabondance des manifestations scientifiques nous a conduits à modifier la périodicité des JEP qui, d'annuelles sont devenues biennales, et à leur conférer un statut de congrès. Cette évolution a impliqué davantage de rigueur mais aussi de lourdeur qui a fait quelque peu perdre à nos JEP leur caractère de rencontre amicale, informelle et spontanée. C'est pourquoi le Comité organisateur en accord avec le Bureau a proposé que les 19es JEP reviennent à leur esprit d'origine en s'ouvrant le plus largement possible aux jeunes chercheurs et à l'exposé des travaux en cours.

Nous ne nous attendions pas à recevoir une réponse aussi enthousiaste à notre appel et le succès a quelque peu noyé le Comité organisateur sous une avalanche de plus de 150 propositions de communications. La sélection ne fut pas une mince opération et devoir caser une centaine de communications en trois jours et demi a soulevé quelques difficultés dont vous ressentirez les effets d'autant que le Groupe répugne aux sessions en parallèle expérimentées à Bruxelles en 1984.

Mais de ce débordement inattendu, il faut tirer des leçons qui, elles, sont hautement positives. Notre communauté s'est largement renouvelée, le nombre de jeunes chercheurs a crû, la qualité et l'éventail des contributions sont frappants. Le potentiel de la recherche en communication parlée a donc été préservé et enrichi. Il faut par ailleurs souligner combien le phénomène de transdisciplinarité que nous avions mis en valeur dès 1970 s'est amplifié au point qu'il est devenu de plus en plus difficile de classer les communications dans les rubriques habituelles de nos rencontres. Ce mouvement d'osmose et de symbiose illustre aussi la complexité des phénomènes qui interagissent au sein de la communication parlée et de la communication homme-machine. La nécessité de recourir de plus en plus aux relations qui unissent les mécanismes centraux aux

processus dits périphériques si elle autorise certains à proclamer l'unicité d'une discipline émergente - par dessus l'éclatement de la phonétique - indique aussi par la même occasion que les rapports d'inclusion avec les disciplines voisines sont fragilisés. Il est significatif que notre réunion ne comporte pas de session consacrée à la neurophysiologie de la parole ou aux neuro-sciences. Des forces centripètes et centrifuges sont manifestement à l'oeuvre dans toutes nos disciplines et ces forces s'exacerbent sous la pression de la R-D. technologique. Il est probable que de nouveaux équilibres vont devoir s'instaurer entre les différents acteurs de la recherche fondamentale et de la recherche industrielle. Cette dernière a déjà adopté d'autres formes de collaboration depuis quelques années grâce notamment aux programmes ESPRIT et à l'implantation des réseaux européens. Au-delà des aspects bénéfiques de la collaboration entre les laboratoires européens, l'intégration européenne en tant que telle soulève d'autres enjeux, de nature culturelle cette fois. Dans le cadre de ces mutations, notre Groupe doit s'interroger sur son devenir qui conditionne aussi son identité.

Si nous avons choisi à l'origine comme cadre d'accueil le Groupement des Acousticiens de Langue Française c'est parce que le GALF réunissait la triple condition de niche scientifique, de communauté culturelle francophone et de support administratif. Le mouvement qui aujourd'hui tente de regrouper au niveau européen les sociétés savantes de physique, de chimie, etc., la création d'EURASIP et d'ESCA, les efforts en vue de fonder un IEEE européen divergent de la notion de confédération de groupes nationaux telle que FASE. Au milieu de ces mouvements et de ces tensions, notre Groupe risque de se diluer et d'égarer une identité francophone dont l'enjeu politique et culturel dépasse, et de très loin, la défense de la langue française en tant qu'instrument de diffusion des connaissances scientifiques. La francophonie du Nord a, surtout ne l'oublions pas, une responsabilité écrasante à l'égard de la francophonie du Sud. Ce n'est pas le moindre des paradoxes que ce soit le rédacteur-en-chef de la revue "*Speech Communication*" qui adjure le Groupe de la Communication parlée de veiller très précisément à la forme que prendra son avenir dans le champ géo-politique instable que nous connaissons.

Max WAJSKOP

ALLOCUTION DE MADAME LE PROFESSEUR F. THYS, RECTEUR DE L'UNIVERSITE

Monsieur le Représentant du Ministre-Président,
Monsieur le Délégué de la Communauté française,

Mesdames et Messieurs,

C'est un double plaisir pour le Recteur de cette Université d'accueillir les 19es Journées d'Etude sur la Parole, l'année même du 30e anniversaire de la création de notre Institut de Phonétique, institut qui a joué dans la fondation du Groupe de la Communication un rôle essentiel.
C'est en effet au cours d'un colloque organisé à l'ULB en février 1968 avec des représentants des universités de Londres, d'Aix, Grenoble et du CNRS que fut prise la décision de regrouper les laboratoires francophones oeuvrant dans les divers domaines de la communication parlée. Cette décision fut concrétisée en février 1970 par les 1ères Journées d'Etude sur la Parole qui eurent lieu à Grenoble.

A cette époque, il n'était pas encore trivial de rassembler les chercheurs de la communauté parole sous cette appellation de "Communication parlée". Les initiateurs de cette création avaient déjà pris conscience non seulement des aspects multidisciplinaires d'un tel regroupement mais dès le départ, ils avaient posé la nécessité de croisements transdisciplinaires tant sur le plan des recherches elles-mêmes que de leurs développements technologiques et industriels.

La Communication parlée avec ses quatre composantes principales, la production de la parole, sa transmission, sa perception et sa compréhension offre une voie royale d'accès aux processus qui gouvernent ce complexe de systèmes que forme l'ensemble intégré des systèmes nerveux, musculaires et sensoriels qui médient l'interprétation cognitive du monde physique et social qui nous environne.

Voie royale ne signifie pas voie aisée. Bien au contraire ! Ayant lu par curiosité les programmes de 1973 et de 1984, je me suis aperçue que celui de 1992 reprenait à peu près les mêmes thèmes. Les disciplines rassemblées dans cette salle et qui vont de la phonologie à la psychologie en englobant toutes les technologies de l'information suffiraient à elles seules à démontrer que la Communication parlée réclame bien davantage que la simple addition de disciplines scientifiques. Chacune d'entre elles a dû se mesurer aux autres, pénétrer ses modus operandi et s'approprier son langage. Cette structuration dynamique croisée d'univers épistémologiques particuliers est depuis longtemps le trait distinctif de votre groupe.

Bien sûr, d'autres domaines regroupés agissent aujourd'hui de même et les historiens noteront que les années 80 et 90 auront vu l'émergence de nouvelles formes de travail scientifique qui sont, en tant que telles, la manifestation que nous sommes définitivement entrés dans l'ère des systèmes complexes et que cette approche est désormais la seule qui nous permette de répondre aux questions que se posent désormais les chercheurs.

Pour ma part, revenant à mes démons familiers d'économiste et de défenseur du concept même d'Université, je souhaiterais relever deux points qui nous concernent tous.

L'informatisation de notre société marque la révolution de l'intelligence et la conquête des esprits, comme la domestication de la terre fut le résultat de la révolution agricole, et comme la révolution industrielle s'accompagna de l'émergence du capital. Or, la conquête des esprits qui est une démarche culturelle apparaît aujourd'hui comme l'instrument le plus efficace d'une politique industrielle et commerciale.

Dans cette politique, les industries de la langue dont vous êtes les artisans précieux forment un enjeu capital non seulement pour la langue française mais aussi pour le développement multilingue harmonieux de l'Europe. Le chiffre d'affaires aujourd'hui consacré aux industries de la langue dans le monde est considérable et appelé encore à s'accroître mais que l'on n'oublie pas surtout, tant dans les sphères nationales qu'européennes, que la recherche industrielle sans le ferment de la recherche fondamentale n'aurait pas l'impact qu'elle a aujourd'hui. On sait depuis les travaux économétriques de ces dix dernières années (cf. JAFFE, 1989 [1]) que le coefficient multiplicateur de la RF sur la RI est proche de 4. C'est dire qu'un franc pour la recherche universitaire dans un état générerait 4F de R.I. dans ce même état. L'efficacité macroéconomique de la recherche universitaire a ainsi été démontrée. Il appartient aux pouvoirs publics - dans le sens le plus large de ce terme - de ne pas négliger cette donnée fondamentale dont le poids culturel et éducatif est encore plus prononcé lorsqu'il s'agit de vos domaines de recherche.

Ces derniers mois ont consacré aussi la volonté de l'U.L.B. de consolider sa réflexion sur le racisme. La parole joue un rôle essentiel dans la question éthique et morale de transmission des valeurs aux générations suivantes. L'Université doit souligner combien sont fallacieux les arguments racistes et rappeler sans cesse leurs conséquences inacceptables.

---

[1] JAFFE, A.B., 1989 : "Real Effects of Academic Research", The American Economic Review, December, nr. 5, vol. 79.

Je souhaite rappeler que dès 1834 les fondateurs de l'Université juraient d'apprendre à leurs élèves "à consacrer leur pensée, leurs travaux, leurs talents au bonheur et à l'amélioration de leurs concitoyens et de l'humanité".

Je vous souhaite de poursuivre avec fruit durant ces 19es Journées d'Etude sur la Parole, le dialogue que vous avez entamé depuis 1970 d'abord entre vous, ensuite avec la communauté scientifique internationale.

19es J.E.P. - Bruxelles, 19 au 22 mai 1992.

ETUDES STATISTIQUES DES MICROPERTURBATIONS DE LA PERIODE FONDAMENTALE

R. DE GUCHTENEERE ET J. SCHOENTGEN*

INSTITUT DE PHONÉTIQUE
UNIVERSITÉ LIBRE DE BRUXELLES

* Fonds National de la Recherche Scientifique, Belgique

Résumé

Nous avons récemment présenté un algorithme de mesure des microperturbations de la période fondamentale du signal de parole. On désigne par microperturbations les légères fluctuations des durées des cycles glottiques. Ces fluctuations sont en général calculées par rapport à une moyenne courante. Cette démarche ignore l'aspect de série temporelle des séquences de périodes (c'est-à-dire le fait que chaque période doit être considérée comme un événement se produisant à un instant précis et non pas indépendamment des périodes précédentes et suivantes). Dans cet article, nous traitons statistiquement les séquences de périodes comme des séries chronologiques. Nous proposons une modélisation à l'aide d'un modèle autorégressif d'ordre peu élevé.

INTRODUCTION

Notre article décrit une nouvelle approche de l'analyse des microperturbations du cycle glottique.

Il est connu que même pendant l'émission de voyelles soutenues, le signal de parole n'est pas parfaitement périodique. Il existe des variations de la durée des périodes, ainsi que de leur amplitude. On désigne ces fluctuations par le terme de microperturbations de la période fondamentale (angl. : jitter) ou de l'amplitude (angl.: shimmer). Dans le cas de locuteurs en bonne santé, ces microfluctuations sont faibles - de l'ordre de 0.1 % de la période fondamentale pendant la phonation soutenue. Elles peuvent croître de façon considérable dans le cas de certaines pathologies laryngées.

Les microperturbations ont été étudiées depuis une trentaine d'années [3]. Par convention, leur ampleur est mesurée en extrayant les durées des périodes successives pendant la phonation soutenue et en calculant un indice de dispersion des durées obtenues sur un laps de temps équivalent à une cinquantaine de périodes. Souvent, les différences prises en compte sont non pas celles entre périodes successives, mais bien les différences entre les périodes individuelles et une moyenne courante. Cette opération a pour but de supprimer les effets à long terme de l'intonation.

En effectuant le type de mesure qui précède, trois hypothèses sont posées implicitement :

1) La première est que, après suppression de toute variation intonative, les différences entre périodes successives sont statistiquement indépendantes. C'est seulement sous cette hypothèse qu'une simple mesure de dispersion suffit à décrire les microperturbations.

2) La seconde est qu'il est possible de supprimer toute variation intonative par une moyenne courante

3) la troisième est que cette opération est licite : ce que l'on supprime ainsi n'aurait rien à voir avec les microperturbations.

Ces hypothèses nous paraissent trop restrictives : elles ne tiennent

absolument pas compte de l'aspect chronologique des séquences de périodes. En effet, chaque période est un événement se produisant à un instant précis, et non pas indépendamment des périodes précédentes et suivantes.

Les premiers résultats que nous avons obtenus montrent que ces hypothèses ne sont pas justifiées [2]. En effet, les différences entre périodes voisines ne sont pas statistiquement indépendantes, mais sont en général fortement corrélées, et les perturbations à court terme paraissent se superposer à d'autres variations stochastiques à moyen terme, qui ne semblent pas être la conséquence d'une variation intonative. Ces observations se retrouvent chez une grande majorité de nos locuteurs.

Nous avons mis au point un programme d'analyse des micro-perturbations qui mesure avec une grande précision les périodes individuelles. Il est suivi d'une batterie de tests statistiques visant à vérifier l'indépendance des périodes consécutives, et à modéliser les corrélations entre périodes voisines si l'hypothèse d'indépendance n'est pas vérifiée.

METHODE

Pour cette étude, nous avons analysé 141 signaux, produites par 47 locuteurs. Les locuteurs devaient soutenir trois voyelles ([a], [i], [u]), à un niveau et une fréquence confortables, de la manière la plus stable possible. Les signaux ont été enregistrés dans une cabine insonorisée. Le microphone était placé approximativement à 5 centimètres des lèvres, fixé à la tête. On a enregistré simultanément le laryngogramme (ou EGG pour électroglottogramme : c'est un signal qui varie proportionnellement à la surface de contact des cordes vocales). Les signaux étaient numérisés à 44 kHz par un processeur audio SONY et enregistrés sur bande vidéo. Ils étaient ensuite numérisés à nouveau à 20 kHz pour être stockés sur le disque dur d'un ordinateur Masscomp 5050.

Notre programme d'analyse utilise le suréchantillonnage pour atteindre une résolution suffisante. Il se déroule en deux étapes principales :

1) Une détection grossière des événements dans le signal est d'abord effectuée. On comprend par événement un passage à zéro dans le signal acoustique, ou un pic dans la dérivée du laryngogramme. Ces événements ont lieu aux environs de l'instant de fermeture des cordes vocales. Une première visualisation des périodes ainsi mesurées permet de repérer des erreurs grossières dans la détection des événements.

2) Ensuite, une portion du signal centrée sur ces événements est suréchantillonnée, puis filtrée, afin d'accroître la résolution temporelle. Nous employons un suréchantillonnage du signal d'un facteur 8, ce qui mène à une fréquence d'échantillonnage effective de 160 kHz.

Le suréchantillonnage permet d'obtenir une résolution temporelle importante, en l'occurrence 6.25 microsecondes, tout en limitant la fréquence d'échantillonnage à une valeur raisonnable, pour faciliter l'acquisition, le stockage et le traitement grossier des signaux.

Après le suréchantillonnage, il convient de vérifier la qualité de l'interpolation effectuée. En effet, le suréchantillonnage ne fait qu'estimer les positions des événements recherchés, les véritables valeurs étant inconnues. W. Hess (1987) propose le critère suivant : les positions brutes, ou positions détectées dans le signal avant suréchantillonnage n'indiquent que peu précisément la position réelle de l'événement recherché. Cette position peut être n'importe où au voisinage de la position brute. Les distances en nombre d'échantillons entre la position d'événement retenue après suréchantillonnage et l'échantillon précédent (avant suréchantillonnage) devraient donc se répartir suivant une distribution uniforme. Nous avons implémenté un test de chi carré afin de vérifier l'uniformité de cette distribution (Fig. 1)

L'algorithme se déroule de la manière suivante : (Fig. 2)

(a) Lecture de l'EGG et du signal acoustique sur disque.

(b) Calcul de la fonction d'autocorrélation de l'EGG afin d'estimer la période fondamentale moyenne.
(c) Filtrage passe-bas de l'EGG et filtrage passe-bande du signal acoustique. Le filtrage de l'EGG est optionnel.
(d) Dérivation de l'EGG.
(e) Détection grossière des pics dans l'EGG et des passages à zéro dans le signal acoustique.
(f) Vérification visuelle des événements détectés et retour optionnel en (e) (les paramètres d'analyse tels que l'intervalle de recherche, l'amplitude minimale des pics ou de leur pente, etc... peuvent être adaptés).
(g) Suréchantillonnage et interpolation du signal autour des événements retenus et localisation fine des événements.
(h) Mesure des durées qui séparent deux événements consécutifs.
(h) Visualisation et tests statistiques.

L'algorithme traite simultanément le signal acoustique et le laryngogramme, sauf si l'un des signaux est trop faible ou trop bruité. L'accord entre ces deux signaux, physiquement fort différents, est une indication de la fiabilité des mesures. L'algorithme offre aussi des possibilités de visualisation graphique des résultats (séquences des périodes, distributions statistiques, perturbations, etc...) et propose une batterie de tests statistiques, destinés notamment à vérifier l'indépendance des périodes consécutives.

TRAITEMENT STATISTIQUE

A l'examen des différents signaux, il apparaît que les périodes individuelles consécutives sont en général fortement corrélées. Ce fait est important, car il mène à reconsidérer les méthodes d'étude des perturbations. Il est évident qu'une mesure de dispersion ne permet pas de tenir compte de cette corrélation de période à période. Nous avons donc utilisé les outils classiques d'étude des séries temporelles, et, après étude des autocorrélations des séquences de période, nous en sommes arrivés à une représentation de ces séquences par un modèle autorégressif d'ordre peu élevé.

RESULTATS ET DISCUSSION

L'algorithme a été testé sur plusieurs centaine de signaux, et ses performances sont très satisfaisantes [2], [7], [8].

Un exemple de modélisation AR est montré à la figure 3. La figure 3a représente la séquence de périodes d'une voyelle [a] produite par un locuteur masculin pendant une seconde. L'analyse de cette séquence, après avoir retranché des périodes individuelles la valeur moyenne de la séquence, suggère un modèle AR(1) de la forme

$$P(n) = 0.89\ P(n-1) + z(n)$$

ou z(n) est un bruit blanc gaussien de moyenne nulle et de variance 0.08. Ce modèle a été utilisé pour synthétiser la séquence de la figure 3b.

Pour apprécier la qualité de la modélisation, nous filtrons le signal avec le même modèle, et nous effectuons un test de chi carré sur les coefficients d'autocorrélation du résidu afin de vérifier qu'il peut être assimilé à du bruit blanc [4]. Des 141 signaux que nous avons étudiés, plus de 120 ont pu être modélisés avec succès (suivant le critère du bruit blanc) par un modèle autorégressif d'ordre 6 ou inférieur. La même analyse a été appliquée avec le même succès à l'étude des microperturbations de l'amplitude.

Les mécanismes sous-jacents à la production des microfluctuations ne sont pas encore complètement compris. Les facteurs neurologiques et cardiaques, qui, on l'a montré [1], [5], contribuent aux fluctuations, devraient produire des perturbations s'étendant sur plusieurs périodes. En effet, en énumérant les mécanismes possibles de production des fluctuations, Pinto et Titze [6] distinguent les contributions à court terme de celles à long terme. Parmi les premières, ils incluent l'asymétrie des cordes vocales, comme par exemple la distribution irrégulière de mucus sur les cordes vocales, ou les différences de masse dues à des pathologies, les turbulences dans la glotte et le couplage entre la glotte et le conduit. Parmi les secondes, ils comptent les facteurs neurologiques. Cette liste suggère l'existence de deux échelles de

temps distinctes, au niveau desquelles différents facteurs agiraient.

Ce que nos résultats suggèrent par contre, c'est l'existence de modèles statistiques qui décrivent les séries chronologiques des durées des périodes en terme d'une composante déterministe, qui filtre un signal purement aléatoire. La distinction entre fluctuations à court terme et à long terme ne s'avère donc pas nécessaire.

REFERENCES

[1] T. Baer (1980), "Vocal jitter : A neuromuscular explanation", Transcripts of the Eighth Symposium of the Care of the Professional Voice, Voice Foundation, New-York, pp 19-22.

[2] R. De Guchteneere and J. Schoentgen (1991), "Mean-term perturbations of the pseudo-period of the glottal waveform", proceedings of the XIIth International Congress of Phonetics Sciences, Aix-en-Provence, pp 354-357

[3] P. Lieberman (1963), "Some acoustic measures of the fundamental periodicity of normal and pathologic larynges", J. Acoust. Soc. Am., 35, pp 344-353

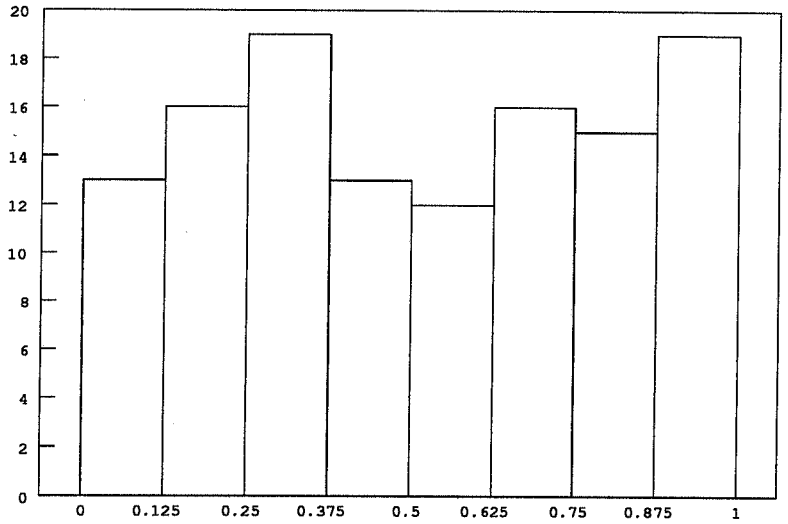
[4] G. Mélard (1990), "Méthodes de prévision à court terme", Editions de l'Université de Bruxelles, Bruxelles, pp 284-291

[5] R.F. Orlikoff and R.J. Baken (1989), "Fundamental frequency modulation by the heartbeat: preliminary results and possible mechanisms", J. Acoust. Soc. Am., 85, pp 888-893.

[6] N.B. Pinto and I.R. Titze (1990), "Unification of perturbation measures in speech signals", J. Acoust. Soc. Am., 87, 3 , pp 1278-1289.

[7] J. Schoentgen and R. De Guchteneere (1991), "An algorithm for the measurement of jitter", Speech Comm., 10, pp 533-538

[8] J. Schoentgen et R. De Guchteneere (1991), "Analyse des microfluctuations des cycles glottiques", Actes du séminaire "Traitement et représentation du signal de parole", Le Mans, pp 108-111



distances entre positions des événements et l'échantillon précédent (avant suréchantillonnage), mesurées en fraction de pas d'échantillonnage

Figure 1

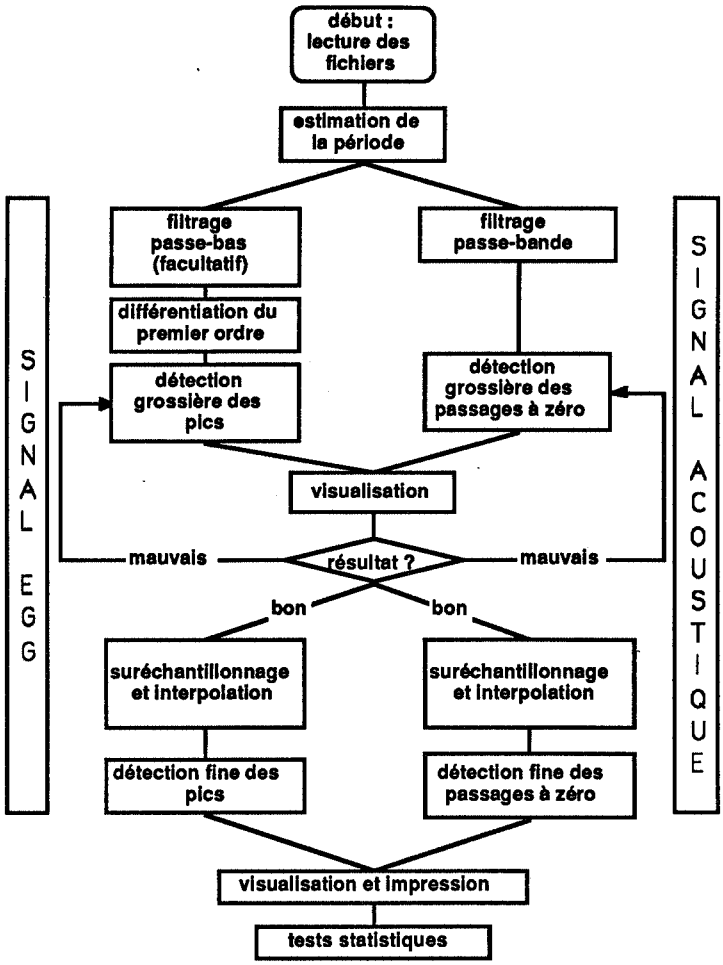
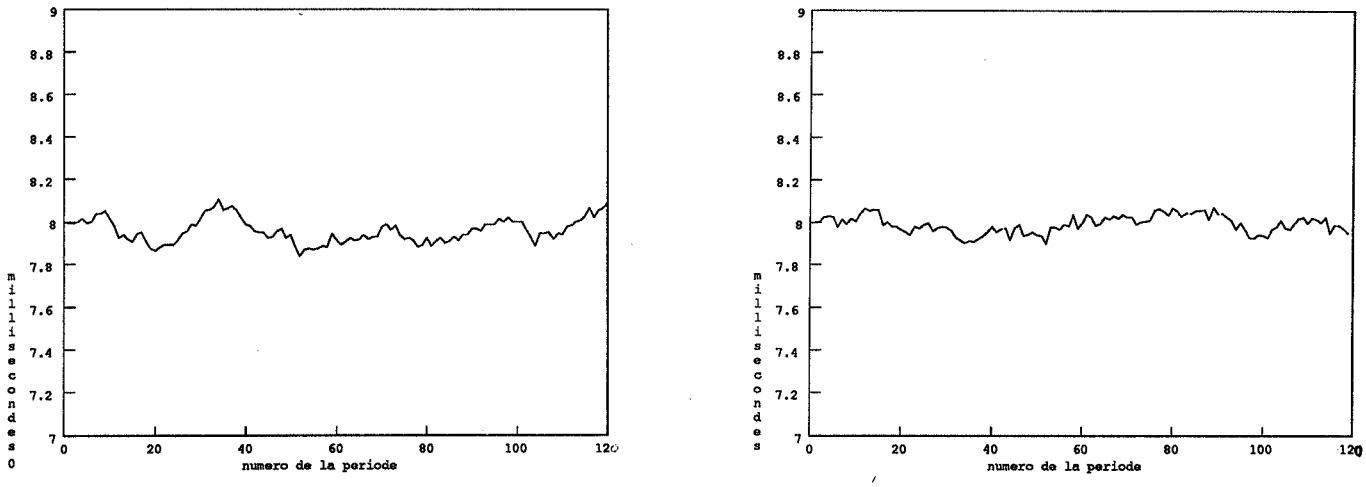


Figure 2



3a 3b

Figure 3

19es J.E.P. - Bruxelles, 19 au 22 mai 1992.

L'IMPORTANCE DES PROCESSUS AERODYNAMIQUES DANS LA PRODUCTION DE LA PAROLE

CELIA SCULLY

DEPARTMENT OF PSYCHOLOGY, UNIVERSITY OF LEEDS

Résumé

Un locuteur français a produit des séquences de sons fricatifs dans deux contextes vocaliques différents, [a..a] et [i..i]. Des données électropalatographiques et aérodynamiques avec une trace A qui indique l'aire d'une constriction du conduit vocal sont analysées, surtout pour les fricatives non-voisées [h], [ç] et [S] ([ʃ]). Pour ces fricatives, la valeur de A, qui indique l'aire minimum d'une constriction du conduit vocal, est moins grande pour le contexte [i..i] que pour celui de [a..a] en général. Les valeurs calculées pour des constrictions composées indiquent que les changements aérodynamiques pour [S] et [ç] dans les deux contextes vocaliques peuvent résulter d'une action constrictive relativement constante, liée à une action pour la voyelle. Pour [h] la chute de pression est distribuée par le locuteur pour régler la distribution des sources acoustiques.

1. INTRODUCTION

Les processus aérodynamiques produisent des liens entre les actions des locuteurs et le signal de la parole. Une des raisons de la complexité des traits acoustiques en parole naturelle est que le système aérodynamique est integré. Les articulations sont quasi-indépendentes l'une de l'autre. Par contre, pendant l'étape aérodynamique les articulations du larynx influencent les sons engendrés par une constriction du conduit vocal; et les changements de la forme de la constriction du conduit vocal produisent un effet sur la vibration des cordes vocales et la source vocale. Puisqu'ils sont liées par la pression orale et le débit volumique, les différentes sources acoustiques montrent des covariations. Comme la pression et le débit dépendent en partie de la forme du conduit vocal, les changements de fréquence et de bande passante des formants montrent aussi des covariations.

Le locuteur connait bien ses propres systèmes de production de la parole. Il doit contrôler les processus aérodynamiques pour arriver à la séquence desirée de sons. Dans certains cas il est necessaire d'éviter une augmentation trop importante de la pression orale. En sus de la source vocale, le locuteur doit contrôler les niveaux des sources de bruit turbulent: bruit 'aspiration' au dessus de la glotte et bruit 'frication' en aval d'une constriction du conduit vocal. Ces deux sources de bruit sont générees pendant la production des sons fricatifs.

Dans notre modélisation nous suivons Stevens (1971):

pour le bruit de frication

$$p_{fric} \propto PC^{1.5}\, AC^{0.5} \qquad (1)$$

et pour le bruit d'aspiration

$$p_{asp} \propto PDIFF^{1.5}\, AG^{0.5} \qquad (2)$$

Le système aérodynamique ressemble en effet à un circuit électrique. Le débit volumique correspond au courant électrique; la pression au voltage; les aires

des constrictions du conduit respiratoire correspondent aux conductances (1/R) électriques. Il y a une différence entre les deux systems integrées: dans un circuit électrique:

$$R = V/I \qquad (3)$$

Ici, dans les processus aérodynamiques de la production de la parole, nous avons:

$$A^2 = k\ 0.5\rho\ U^2/P \qquad (4)$$

A est l'aire minimum d'une constriction, U est le débit volumique d'air, P est la chute de pression a travers la constriction, ρ est la densité de l'air et k est une constante empirique. Dans la geometrie compliquée et changeante du conduit respiratoire la valeur de k n'est pas bien connue, mais nous pouvons choisir une valeur de 1, pour simplifier (Scully, 1986), ce qui donne l'expression:

$$AC = 0.00076\ UC/PC^{0.5} \qquad (5)$$

pour une constriction du conduit vocal et

$$AG = 0.00076\ UG/PDIFF^{0.5} \qquad (6)$$

pour la glotte, avec A en cm^2, U en cm^3/s et PC ou PDIFF en cmH_2O. PC est la chute de pression à travers une constriction orale, PDIFF est la chute de pression à travers la glotte, PSG-PC.

En général, les débits UG et UC n'ont pas la même valeur. Ils diffèrent si la pression orale augmente ou diminue, et aussi s'il y a un changement de VC, le volume de la cavité du conduit vocal entre la glotte et la constriction orale (voir Figure 1). Mais si on examine des périodes près des pics de PC et au milieu des périodes de voyelles où la mâchoire ne bouge pas trop, on peux considérer que ces deux débits volumiques UG et UC ont la même valeur U.

Dans ce cas, on peut mettre ensemble les deux constrictions qui sont en série et exprimer leurs effets conjoints par une constriction composée ACP, où:

$$1/(ACP)^2 = 1/(AG)^2 + 1/(AC)^2 \qquad (7)$$

Le débit volumique U dépend de cette constriction effective ACP, si la pression subglotique PSG reste constante, parce que:

$$ACP = 0.00076\ U/PSG^{0.5} \qquad (8)$$

2. LES EFFETS PRODUITS PAR L'AUGMENTATION DE L'AIRE DE LA GLOTTE

Si la glotte est élargie, ACP devient plus grande et le débit volumique U augmente aussi. Il en résulte que la pression orale monte plus vite. Dans la production de fricatives non-voisées cette ouverture de la glotte est une action essentielle pour donner une source fricative bien forte (formule 1) et pour faire cesser la source de la voix. En même temps, la source aspiration est augmentée (formule 7).

Ainsi, une seule action du larynx produit des changements dans les trois sources.

Dans notre modélisation nous avons montré que si la glotte augmente trop en avance dans la production du mot anglais "polite", le mot est realisé comme [pəˈlaiçt] avec une fricative palatale supplémentaire, un resultat non desiré pour l'anglais (Scully, 1987, p.137).

Ici on va décrire les actions d'un locuteur français dans la production de quelques fricatives non-voisées et dans deux contextes différents [a..a] et [i..i]. On va analyser surtout [h], [ç] et [S] ([ʃ] comme dans le mot "chou"). La fricative [ç] n'est pas un son français, et le locuteur a choisi une façon de production intéressante: il a gardé la langue assez immobile entre [i] et [ç].

Dans la production de [S] il est prévu que le locuteur va former une constriction entre la pointe et le dos de la langue et le palais. Pour [h] il est prévu que le conduit vocal ne changera pas entre [i] et [h]. Mais celui-ci est un cas où il est nécessaire d'éviter une trop grande augmentation de la pression orale (Scully, 1987, p.123).

3. DONNEES POUR LE LOCUTEUR PB

Le locuteur PB est un homme français, un des deux locuteurs pour un projet européen.

3.1 MESURES AERODYNAMIQUES

Figure 2 montre les traces pour [S], [ç] et [h] dans deux contextes [a..a] et [i..i], y compris le débit, la pression orale, un paramètre appelé A qui indique AC (formule 5), et des traces acoustiques. Les méthodes ont déjà été décrites (Scully, 1986; Scully et al., 1992). Il faut souligner que cette trace A n'est pas la véritable aire minimum de la constriction du conduit vocal. Le signal A ($=U/P^{0.5}$) est obtenu des signaux U et P (Aerodynamic Speech Analyser, R.Caley). Pour des fricatives différentes on tente de mesurer les paramètres indiquées dans la Figure 2. A med est la valeur de cette paramètre A au milieu de la constriction fricative. Figure 3 montre les valeurs de cette mesure pour les fricatives non-voisées du locuteur PB. On voit que la valeur A est moins grande pour le contexte [i..i] que pour celui de [a..a] en général. Il n'est pas toujours possible de mesurer A med. Dans ces cas on mesure A au milieu du bruit frication-aspiration indiqué par une des traces acoustiques. En chaque cas la mesure est près de la valeur minimum pour A. Pour [x] la pression orale n'est pas obtenue par le tube orale. Dans ce cas il est nécessaire d'éstimer la pression orale par la pression sublottique. Celle-ci dans son tour est estimée par les pics de pression pour [p] avec interpolation entre le [p] à chaque côté.

Les traces aérodynamiques indiquent que le locuteur produit une voyelle [i] assez avancée et assez haute auprès de la voyelle [i] de l'anglais par example. Pour la voyelle [a] de PB la pression orale obtenue par un tube avec son ouverture au dessous du palais est zéro (pression de l'atmosphère). Mais dans sa voyelle [i] la pression orale est augmentée et on peut estimer sa valeur. Avec la valeur du débit on peut estimer l'aire du conduit vocal AC pour [i] (formule 5 et la trace A, Figure 2). La valeur de l'aire de la glotte est estimée de la même manière (formule 6), mais avec estimation de pression subglottique par interpolation entre les valeurs dans le [p] à chaque côté (Scully, 1986).

Les Tables I et II donnent les resultats.

Table I Valeurs estimées pour AC et AG, voyelle [i], cm^2.

	AC	AG
[iSi]	0.13	0.050
[içi]	0.10	0.053
[ihi]	0.14	0.047

La Figure 2 montre (trace A) que la constriction totale du conduit vocal devient un peu plus sévère pour [S] par rapport à [i], mais qu'elle reste constante pendant toute la séquence [içi].

La Table II donne les valeurs obtenues pour AC (l'effet total du conduit vocal devant le tube pour la pression orale) et AG, l'aire de la glotte.

Table II Valeurs estimées pour AC et AG dans les fricatives, cm^2.

	AC	AG
[aSa]	0.20	
[iSi]	0.12	
[aça]	0.25	
[içi]	0.12	
[aha]		0.43
[ihi]	0.33	0.32

3.2 MESURES ELECTRO-PALATOGRAPHIQUES (EPG)

Figure 4 montre des cadres EPG pour les séquences [aSa], [iSi], [aça] et [içi]. (Malheureusement, nous n'avons pas obtenu de données EPG pour [h].) On voit que la constriction est plus sévère (avec un plus grand nombre de contacts) pour les fricatives dans le contexte [i..i] que dans celui de [a..a].

La Figure 4 montre aussi que la configuration de la langue est la même pour les voyelles et pour la fricative dans la séquence [içi].

Pour [iSi] les contacts sont reduits de [i] a [S], c'est

à dire que la constriction est moins sévère pendant [S] que pendant [i]. Les données EPG et aérodynamiques sont d'accord pour [ç], mais pas pour [S]. Le diminution de la valeur de A pour [S] est peut-être attribuable à la configuration des lèvres et des dents.

4.DISCUSSION

4.1 [aça] et [içi]

L'aire de la constriction pour la fricative [ç] est reduite dans le contexte [i..i]. Si on prend les valeurs pour AC [i], 0.10 cm^2, et pour AC [ç] dans le contexte [a..a], 0.25 cm^2, on peut calculer l'effet total pour le conduit vocal (formule 7) qui donne:

$$ACP = 0.09\ cm^2$$

(ACP est la constriction composée pour le conduit vocal).

Cette valeur est assez près de la valeur obtenue pour [ç] dans le contexte [i..i], (0.12 cm^2). C'est comme si les effets de la consonne et de la voyelle produisent un effet additionnel, bien qu'il n'est pas question en réalité de deux constrictions en série. La plus grande partie de la pression subglottique apparaît a travers la constriction composée orale, ce qui doit donner une bonne fricative [ç].

4.2 [aSa] et [iSi]

On voit le même processus dans la production de [aSa] et [iSi]. Pour [S] dans le contexte [i..i], ACP pour le conduit vocal a des composantes AC [aSa], 0.2 cm^2, avec AC [i], 0.13 cm^2, qui donne:

$$ACP = 0.11\ cm^2$$

près de la valeur pour [iSi] dans la parole de PB. Mais ici la région de constriction doit être plus avancée pour la fricative [S] que pour la voyelle [i].

4.3 [aha] et [ihi]

Le locuteur élargit la constriction du conduit vocal pour passer de [i] à [h]. Pourquoi? Si la constriction du conduit vocal restait a sa valeur pour [i], 0.14 cm^2, pendant que l'aire de la glotte augmentait de 0.047 cm^2 a 0.32 cm^2 pour [h], la valeur de la constriction totale pour le conduit respiratoire, avec composantes AG et AC, deviendrait 0.127 cm^2 (formule 7). Le débit deviendrait 423 cm^3/s (formule 8) et la pression orale PC augmenterait a 5.3 cmH_2O; PDIFF serait reduite a 1.1 cmH_2O. Dans ce cas le bruit frication au palais serait trop fort à côté du bruit aspiration glottal pour [h]. Mais comme PB élargit la constriction du conduit vocal à 0.33 cm^2 la pression orale monte à seulement 3.1 cm^2. La chute de pression est distribuée également entre la glotte et le conduit vocal.

5. CONCLUSIONS

Ces exemples de la production de quelques consonnes dans des contextes vocaliques différents indiquent que les changements aérodynamiques pour certaines fricatives peuvent résulter d'une action constrictive pour le conduit vocal qui est relativement constante, indépendente du contexte, liée à une action pour la voyelle.

La chute de pression est distribuée par le locuteur pour régler la distribution des sources acoustiques.

Nous avons représenté les processus pour la source vocale du locuteur PB dans notre modèle de la production de la parole (Allwood & Scully, 1982; Scully, 1986, 1987; Scully et al., 1992). Il nous reste à simuler les séquences analysées ici. Il est prévu que les deux contextes vocaliques [a..a] et [i..i] produiront une multiplicité de traits acoustiques opposés. Les effets sont déjà bien connus par les auditeurs et seront utiles pour améliorer les voix synthétiques.

REMERCIEMENTS

Recherche aidée par le projet européen SCIENCE No. SCI*0147-C(EDB). Je remercie beaucoup le locuteur Pierre Badin, également Eric Brearley qui nous a aidé avec l'acquistion des données.

REFERENCES

Allwood, E. & Scully, C. (1982) A composite model of speech production, *Proc. Intl. Congr. Acoust., Speech & Signal Processing, IEEE,* paper S6.6, vol.2, 932-935.

Scully, C. (1986) Speech production simulated with a functional model of the larynx and the vocal tract, *J. of Phonetics, 14,* 407-414.

Scully, C. (1987) Linguistic units and units of speech production, *Speech Communication, 6,* 77-142.

Scully, C., Castelli, E., Brearley, E. & Shirt, M. (1992) Articulatory paths and aerodynamic patterns for some fricatives, *J. of Phonetics, 20,* 39-51.

Stevens, K.N. (1971) Airflow and turbulence noise for fricative and stop consonants: static considerations, *J. Acoust. Soc. Amer., 50,* 1180-1192.

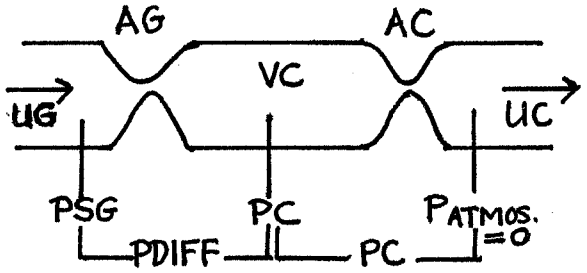


Figure 1. Deux constrictions en série, aire minimum AG et AC, avec débit volumique UG et UC, pression PSG, PC, Patmos (0), volume de la cavité VC.

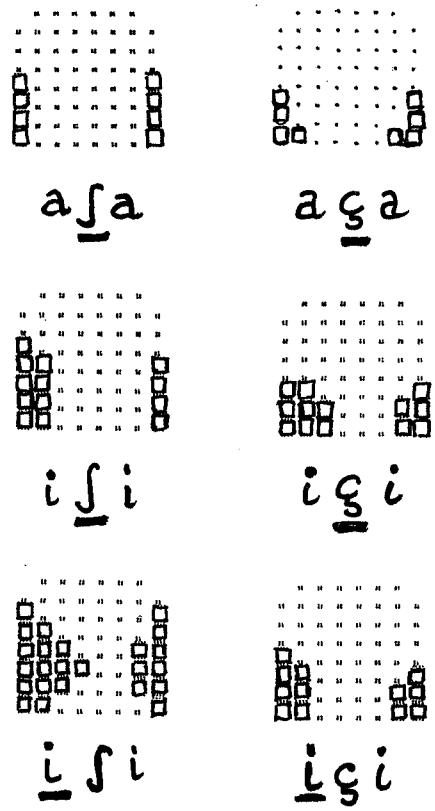


Figure 4. Exemples des cadres EPG pour le locuteur PB: [aSa], [iSi], [aça] et [içi].

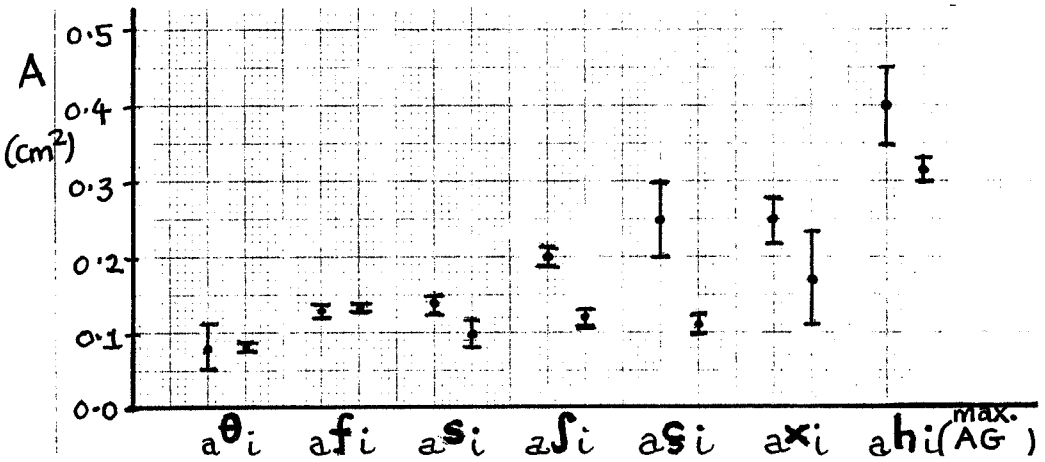


Figure 3. Valeurs pour A (près de la valeur minimum) pour les fricatives non-voisées produites par le locuteur PB. Sept répétitions de chacune. Deux contextes vocaliques [a..a] et [i..i]. Valeurs moyennes et 95% C.I.

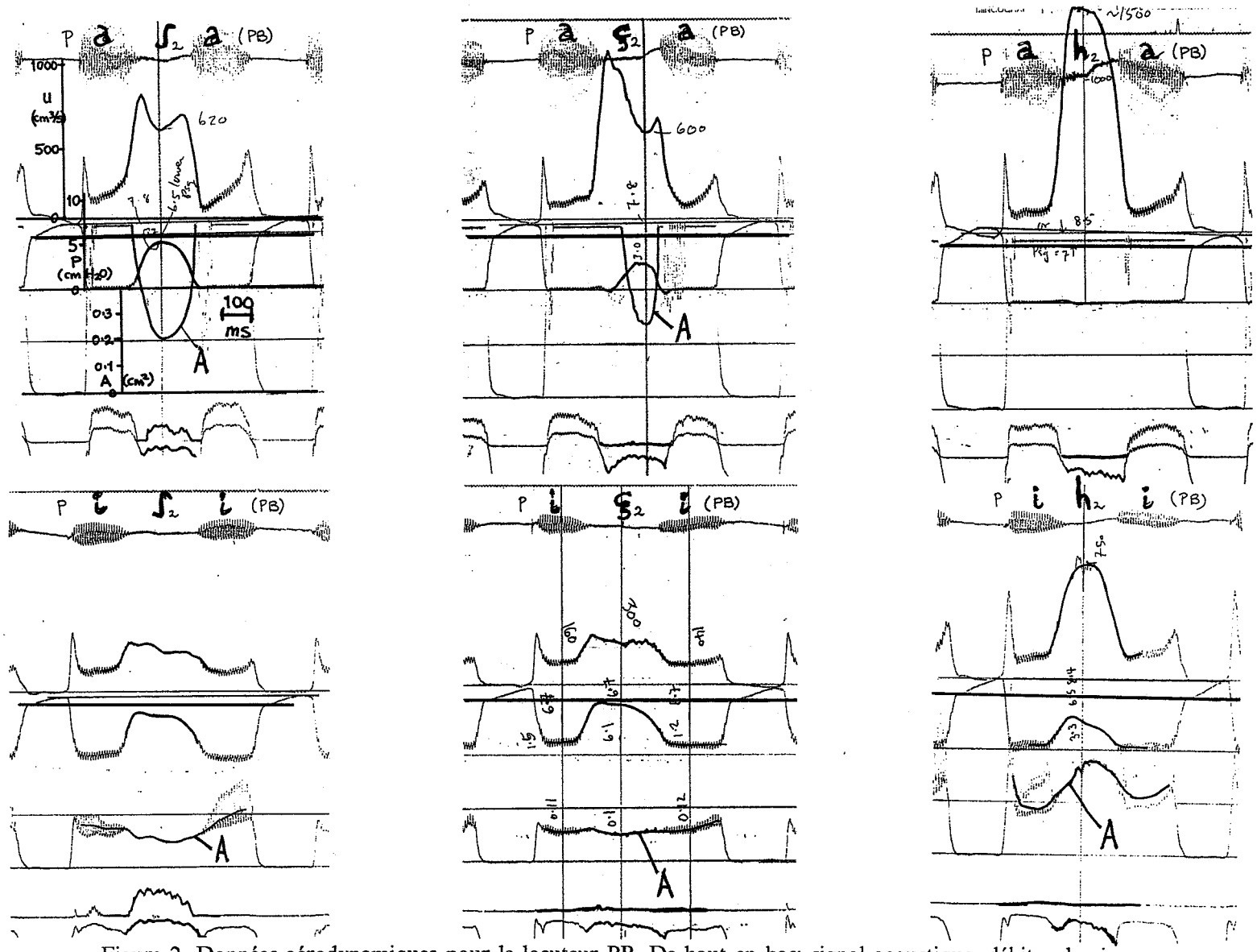


Figure 2. Données aérodynamiques pour le locuteur PB. De haut en bas: signal acoustique, débit volumique (total, pour débit oral), pression orale, A, intensité acoustique hautes fréquences, intensité acoustique fréquences moyennes.

EFFETS DE COUPLAGE SUBGLOTTIQUE : MESURE ET MODELISATION DANS LE DOMAINE FREQUENTIEL POUR LES FRICATIVES D'ARRIERE DE L'ARABE

AMAR DJERADI Laboratoire de la Communication Parlée
Institut Electronique USTHB-El Alia - Bab Ezzouar BP 32 Alger Algerie

PIERRE BADIN & BERNARD GUERIN
Institut Communication Parlée INPG/ENSERG & Université Stendhal
46 Avenue Felix Viallet 38031 Grenoble cedex France

Résumé :

Les effets de couplage des parties subglottiques sur la fonction de transfert du conduit vocal sont évalués grace à une exploration du système phonatoire humain par une séquence pseudo-aléatoire pour les trois conditions de glotte : glotte ouverte, glotte fermée et glotte variable.

Par ailleurs il est possible, lors de la même phase d'expiration correspondant à une configuration articulatoire soutenue, d'enregistrer le son lui-même et le signal utile pour la mesure de la fonction de transfert.

La comparaison des différentes données ainsi obtenues permet la mise en évidence des effets de couplage subglottique d'une part et montre que les conditions à la glotte ont une influence sur les résonances des cavités supraglottiques d'autre part : en effet, nous avons observé une augmentation des fréquences de résonance, des bandes passantes et une modification de l'écart entre les résonances du conduit vocal. Nous avons également noté l'apparition de résonances supplémentaires pouvant correspondre à des formants subglottiques.

Dans une seconde phase, nous avons reproduit ces effets par des simulations sur des configurations géométriques obtenues par la méthode des rayons X sur le même sujet et pour les mêmes configurations.

1 Introduction

Dans cette étude nous présentons les résultats d'une analyse expérimentale et théorique des effets de couplage subglottique sur la fonction de transfert des cavités vocales. Pour cela nous utilisons la méthode développée par DJERADI et al (1991) à l'ICP. Cette méthode est basée sur l'excitation transcutanée du conduit vocal, au niveau de la glotte, par un excitateur délivrant une séquence pseudo-aléatoire. La FFT de l'intercorrélation de ce signal avec le signal capté aux lèvres fournit la fonction de transfert acoustique du conduit vocal.

Notre analyse porte sur les cinq fricatives d'arrière de l'arabe /x, ħ, ʕ, ɣ, h/, pour un sujet homme.

A partir de données radiographiques obtenues sur le même sujet, et pour les mêmes configurations articulatoires, des fonctions d'aire ont été déterminées et les fonctions de transfert correspondantes sont calculées dans les mêmes différentes conditions glottiques (glotte fermée, et glotte ouverte). Pour obtenir une bonne correspondance entre les fréquences des résonances subglottiques mesurées et simulées, il a été nécessaire d'accorder les paramètres du système phonatoire.

Théoriquement le système phonatoire peut être représenté par le schéma suivant :

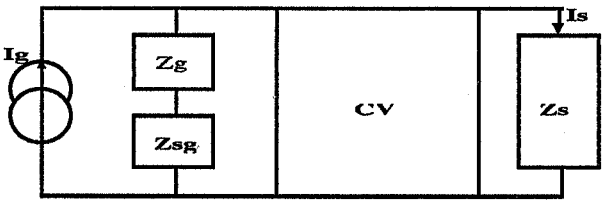


Fig.1. Schéma équivalent du système phonatoire

Ig : débit au niveau de la glotte
Zg : impédance glottique
Zsg : impédance subglottique
CV : quadripôle représentant le conduit vocal
Zs : impédance de sortie (rayonnement aux lèvres)
Is : débit de sortie (mesuré aux lèvres)

La fonction de transfert du système est déterminée par le module du rapport entre le débit aux lèvres et le débit à la glotte :

$$FTR = \left| \left| Is/Ig \right| \right|$$

Cette quantité dépend de l'impédance d'entrée du quadripôle mais aussi des deux grandeurs Zg et Zsg. Dans le cas où la glotte est ouverte, Zg peut être considérée

comme un court circuit et la fonction de transfert du système de phonation dépend de façon importante de l'impédance Zsg. Les caractéristiques acoustiques des parties subglottiques, représentées ici par Zsg, seront alors visibles sur la courbe de réponse du système. Par contre dans le cas où la glotte est fermée, Zg devient infinie et l'impédance Zsg est négligeable devant Zg, la fonction de transfert sera uniquement représentative des cavités supraglottiques. Dans le cas de la phonation, l'aire à la glotte varie avec les vibrations des cordes vocales, on estime alors une ouverture moyenne de la glotte dont l'aire équivalente est égale à la moyenne des aires prises par celle-ci aux différents instants de la phonation. Par conséquent Zg sera aussi variable et aura une valeur moyenne pour laquelle les effets de Zsg pourront apparaître de façon plus ou moins importante sur la fonction de transfert. Ce sont ces différentes situations que nous examinerons au cours de cette étude.

Nous chercherons aussi à déterminer le comportement de ces effets subglottiques face à une modification de la géométrie des cavités supraglottiques, en travaillant sur une opposition de lieu articulatoire "uvulaire/ pharyngale". Nous examinerons aussi l'opposition des deux classes par leur mode articulatoire "voisée/non voisée". L'analyse porte sur 640 enregistrements dont 320 courbes en phonation, 160 courbes à glotte ouverte et 160 à glotte fermée.

2. Mesures expérimentales et analyse des résultats

2.1 Le mode opératoire

La première difficulté à laquelle nous nous sommes confrontés lors de cette étude est le manque de maîtrise dans le contrôle de l'état glottique : en effet il n'est pas facile de garder sa glotte ouverte ou fermée et de maintenir en même temps sa configuration articulatoire constante surtout en mode fricatif et de reproduire de façon stable ces gestes d'une expérience à une autre. Il a fallu donc une longue phase d'apprentissage avant de pouvoir contrôler la contraction glottique.

Pour mesurer une fonction de transfert du conduit vocal à glotte fermée, on procède de la manière suivante : le sujet est placé en chambre sourde, il commence par positionner la source d'excitation pseudo-aléatoire, en plaçant le "vibrateur" sur la peau de son cou, au niveau de sa glotte, pour cela il excite son conduit à l'aide d'un bruit blanc et le retour casque lui permet d'apprécier l'effet acoustique relatif à la position spaciale de cette source, il maintient alors cette position du "vibrateur" et recherche la position correcte de ses articulateurs, en produisant normalement le son à analyser, une fois celle-ci est atteinte, il bloque alors complètement sa respiration tout en maintenant sa configuration constante et c'est à cet instant que l'on effectue la mesure. La détermination de la fonction de transfert à glotte ouverte s'effectue de la même façon qu'à glotte fermée sauf qu'il faut respirer continuement lors de la mesure et essayer par la sensation que procure le passage de l'air sur les articulateurs d'obtenir un contrôle plus précis (du geste) de la constriction sans toutefois prononcer le son. Il faut surtout veiller à ce que la glotte reste constamment ouverte et la configuration géométrique stable pendant toute la durée de la mesure.

2.2 Les effets de couplage subglottiques

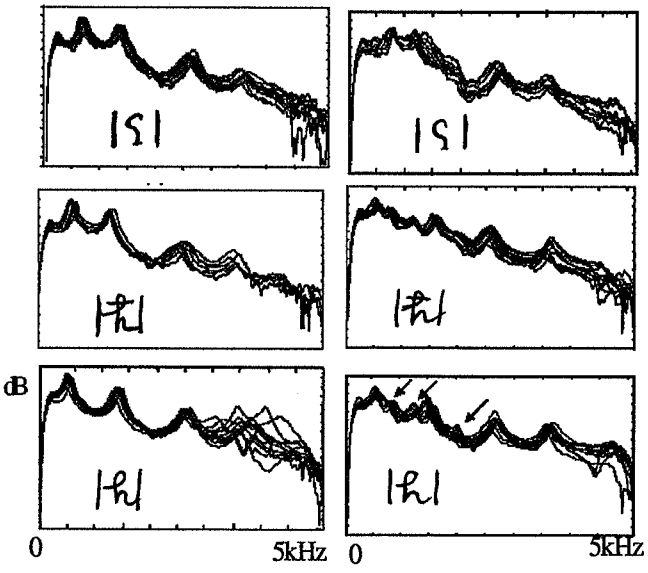
Nous avons pu constater que les fonctions ainsi mesurées, sont assez complexes. Malgrés cette complexité apparente, un grand nombre de pics et de vallées spectraux semble garder des valeurs constantes.

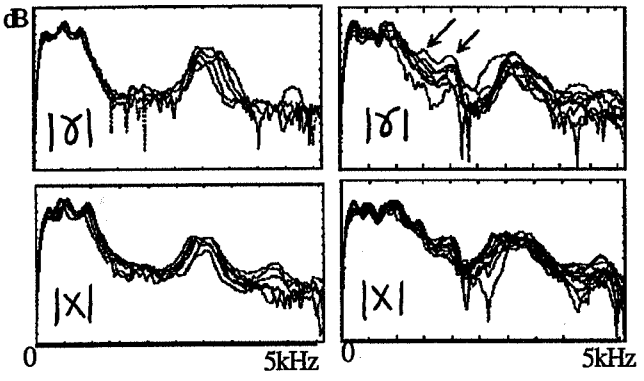
Par ailleurs, les courbes obtenues à glotte ouverte sont proches des courbes mesurées en phonation. Nous observons clairement l'apparition de pics spectraux supplémentaires sur les courbes de réponse à glotte ouverte et qui n'existe pas sur les courbes obtenues à glotte fermée, ces pics ont une amplitude plus faible.

D'autre part l'anti-résonance visible sur les courbes "uvulaires" à glotte ouverte et en phonation vers 2500Hz, disparait sur les courbes à glotte fermée. La présence de pics supplémentaires s'accompagne d'une diminution quasi-systématique de l'amplitude des résonances principales (si l'on compare celle-ci à celle des résonances mesurée à glotte fermée). Sur les courbes de la figure 2, on voit clairement que l'écart de fréquence entre les résonances Fp3 et Fp4 a tendance à s'agrandir avec l'ouverture de la glotte.

On voit clairement que les fricatives possédant le même lieu articulatoire, présentent des fréquences de résonance quasi-identiques, que ce soit pour les deux uvulaires que pour les deux pharyngales. Par conséquent, on en déduit que les formes géométriques sont proches.

La faiblesse des valeurs des écart-types témoigne de la stabilité et de la représentativité de nos mesures sur l'ensemble du corpus.





Fig,2. Fonctions de transfert expérimentales

Par ailleurs nous remarquons une tendance à la hausse des deux premières fréquences et de la quatrième fréquence de résonance et une tendance à la baisse de la troisième fréquence de résonance lorsqu'on passe du rétrécissement "uvulaire" au rétrécissement "pharyngal".

Le passage d'une constriction pharyngale vers une constriction laryngale entraine une descente de la première résonance et une montée de toutes les autres.

Le recul du lieu de la constriction depuis la zone uvulaire jusqu'à la région du haut pharynx a pour effet de diminuer la cavité arrière ce qui se manifeste par une hausse des fréquences de résonance corespondantes, probablement Fp1, Fp2 et Fp4, ceci d'une part, mais d'autre part, le volume de la cavité avant croit ce qui entraine une baisse de la fréquence de résonance correspondante, probablement Fp3. Nous avons relevé les fréquences des pics de résonance sur les fonctions de transfert dans le cas de la glotte ouverte. On retrouve les mêmes tendances dans l'évolution des fréquences avec le recul du lieu de la constriction.

La comparaison des deux types de courbes montre clairement la présence de 3 à 4 pics spectraux supplémentaires, ces pics se situent dans le cas des "uvulaires" aux fréquences 430, 1000, 1400 et 2000Hz et dans le cas des pharyngales dites encore"verticales", ces pics sont visibles aux fréquences 430, 750, 1200 et 2000Hz. On note donc que seulement deux fréquences se retrouvent dans les deux cas, 430 et 2000Hz. On conclue que les autres fréquences dépendent des cavités supraglottiques : en effet lorsque ces fréquences sont voisines(750 et 1200Hz pour les "verticales" et, 1000 et 1400Hz pour les "horizontales"), un seul pic apparait dans une position fréquentielle proche de celle de la résonance supraglottique mais avec une bande passante plus grande. Si nous généralisons cette idée de superposition aux autres fréquences manquantes dans l'une ou l'autre des deux classes de consonnes, nous pouvons conclure en disant que notre sujet présente 6 fréquences de résonance des cavités subglottiques, ce sont : 430, 750, 1000, 1200, 1400, 2000Hz.

2.3 Conclusion

Nous pouvons affirmer que les effets de couplage des parties subglottiques se manifestent sur la fonction de transfert du conduit vocal par la présence de pics supplémentaires. Ces pics sont situés pour notre locuteur aux fréquences : 400, 700, 1000, 1200, 1400 et 2000Hz. La modification de la forme du conduit vocal agit sur ces fréquences de façon légère. Par ailleurs l'amplitude des pics issus du couplage des parties subglottiques et des parties supraglottiques est faible par rapport à celle des pics principaux émanant des cavités supraglottiques.

3. Simulation des effets subglottiques

3.1 Modélisation 3D du conduit vocal

La modélisation des mécanismes de production des fricatives nécessite une connaissance précise des dimensions du conduit vocal. Le choix de la méthode aux rayons X, pour l'obtention des coupes sagittales est intéressant d'une part par le fait que nos objets d'étude sont des fricatives soutenues, et d'autre part par le souci d'obtenir une bonne qualité d'image sans trop exposer le sujet à de fortes doses de rayons X et ceci en admettant que le profil mi-sagittal des fricatives voisées et celui des fricatives non voisées sont très proches (Bothorel & al, 1986). La source d'émission de rayons X est située à 5 ou 6 m du sujet à bombarder et la plaque photosensible, destinée à recevoir la projection de la tête du locuteur, est placée à 5cm de celui-ci. On peut négliger ainsi la distorsion qui résulte de cette projection. Pour obtenir un bon contraste de certaines parties molles de la tête (langue, lèvres..), on utilise des filtres adéquats en aluminium. L'image obtenue est la projection sur un plan parallèle au plan sagittal de toute la tête.

L'expérience consiste à demander au sujet de soutenir la fricative prononcée le plus longtemps possible, ceci afin d'obtenir des sons stables d'une expérience à l'autre.

Dès la stabilisation du son, on fait une acquisition quasi simultanée de l'image aux rayons X, du son et de la photographie des lèvres.

Par ailleurs des moulages en plâtre des parties dures du conduit vocal nous ont permis de mieux préciser les dimensions de cette région.

3.1.1 Acquisition des contours

Cette étape ne peut se faire que manuellement, car il est impossible de déterminer les profils sagittaux du

conduit vocal par un traitement d'image automatique. En effet, la lecture des radiographies n'est pas toujours aisée, il a fallu souvent consulter des spécialistes et discuter la décision qui ne fait pas toujours l'unanimité ; les profils ont donc été tracés manuellement sur des feuilles de nylon transparentes. Nous constatons que le contour des lèvres est très lisible pour l'ensemble des configurations. Le tracé de la partie dure du palais est ajusté grâce à la découpe de cette zone obtenue par moulage en plâtre. Par contre, au niveau du velum, l'interprétation reste très délicate à cause des retours latéraux du velum qui sont plus apparents. La paroi arrière du naso pharynx et la glotte sont facilement repérables pour toutes les configurations. Au niveau de la langue, nous avons pris la décision de suivre le sillon et non pas les rebords supérieurs.

Nous avons aussi relevé quelques repères osseux dans le but d'uniformiser les échelles de mesure des différents éléments géométriques du conduit vocal pour l'ensemble des configurations.

3.1.2 Digitalisation des contours

Dans cette étape nous avons utilisé un système de traitement d'image développé par le laboratoire de traitement d'image et de reconnaissance de formes de l'Institut National Polytechnique de Grenoble (S. Olympieff). L'intérêt de ce système par rapport aux mesures traditionnelles est, outre l'automatisation de la mesure, la grande précision dans la détection des contours, on arrive ainsi à une meilleure estimation de la longueur du conduit vocal et de l'aire aux lèvres. Ce système fournit pour chaque contour, les coordonnées X,Y de tous les pixels (ces pixels sont contigus et ordonnés dans la sens de parcours du contour).

3.1.3 La grille de mesure

L'emploi d'une grille pour déterminer la fonction sagittale est maintenant classique (Heinz & Stevens, 1965 et reprise par Maeda, 1979). La discrétisation du conduit vocal n'est pas uniforme puisque la grille est composée de 3 parties différentes. Entre la glotte et le bas du pharynx, cette grille est constituée de droites parallèles; du bas pharynx à la cavité buccale, on utilise un système de rayons d'une portion de cercle dont le milieu est le point concordant des différentes portions de droites. Cette grille divise donc le conduit vocal en un certain nombre de secteurs, et chaque section du profil sagittal correspond alors à la zone du conduit vocal comprise entre les deux droites qui définissent le secteur.

Des repères fixes sont choisis par rapport à des invariants anatomiques liés au crâne et à la colonne vertébrale du sujet. L'axe vertical du système est choisi parallèle à la paroi du fond du pharynx. Le centre O est le centre du cercle qui approxime le contour du voile du palais. Nos repères (l'incisive et la partie dure du palais) nous ont servi à définir un même encadrement servant d'échelle pour l'ensemble des contours.

3.1.4 Détermination des fonctions sagittales

On arrive ainsi à délimiter les différentes sections par des petits contours fermés dont on calcule le centre de gravité et la surface en pixels. La ligne médiane du conduit vocal est constituée de segments joignant ces centres de gravités. La longueur de chaque section est alors obtenue en mesurant sur cette ligne médiane, la longueur du segment joignant les deux droites de la grille délimitant cette section. La distance sagittale est enfin calculée comme la surface de la section divisée par sa longueur, nous obtenons ainsi les fonctions sagittales fig(4).

3.1.5 Détermination des fonctions d'aire

Le passage de la fonction sagittale à la fonction d'aire se fait par la méthode classique proposée par Heinz & Stevens [1965]. Le principe consiste à appliquer la relation suivante entre l'aire A et la distance sagittale d (c'est à dire la hauteur d'une section) : $A = a.d^b$ où a et b sont les coefficients qui dépendent de la région du conduit vocal.

Les fonctions d'aire initiales des 5 fricatives sont obtenues par l'application des coefficients déterminés par Perrier et al sur les voyelles françaises. Puis à l'aide des fonctions de sensibilité telles qu'elles sont définies par Fant & Pauli (1974) ou Mrayati &Carré (1976) et des distributions de la pression et du débit dans le conduit vocal, nous avons ajusté les fonctions de transfert initiales de façon à obtenir une bonne superposition des fonctions de transfert mesurées avec celles obtenues par simulation .

3.2 Le modèle de simulation des effets subglottiques

Pour simuler les effets de couplage des parties subglottiques sur les parties supraglottiques, nous avons représenté le conduit vocal par sa fonction d'aire et les parties subglottiques par une cellule de Foster à trois résonances (600, 1000, 2000 Hz équivalentes à celles mesurées lors de la phase expérimentale).

Le modèle complet, dont le schéma de principe est donné à la figure 1, permet aussi d'intégrer les différentes pertes telles que : pertes par chaleur, pertes de viscosité et celles dues à l'impédance des parois. Par contre, dans le cas de la glotte ouverte, il ne peut tenir compte de la résistance à la constriction. L'aire à la glotte est notre paramètre variable, sur lequel nous jouons pour approcher au mieux les effets de couplage donnés dans la

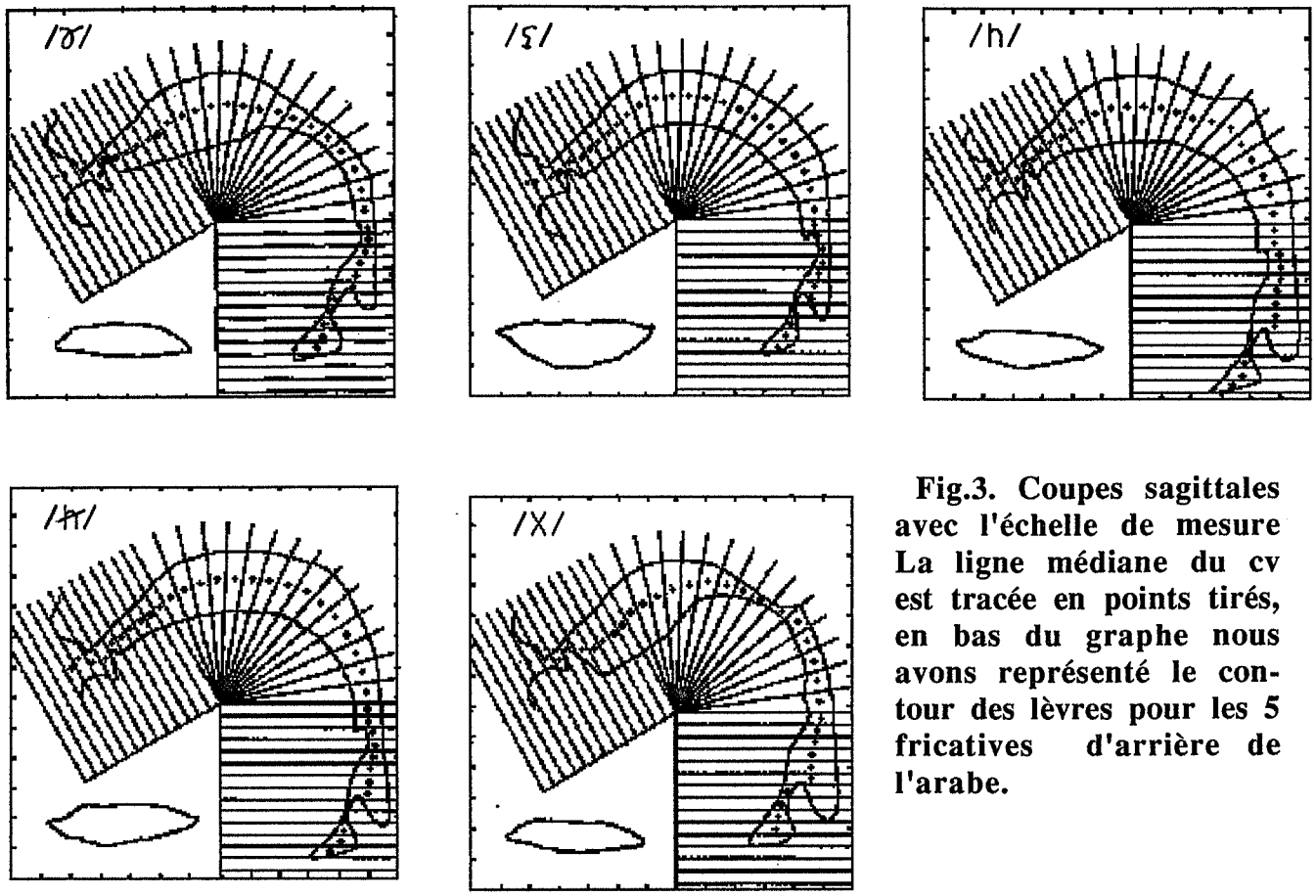
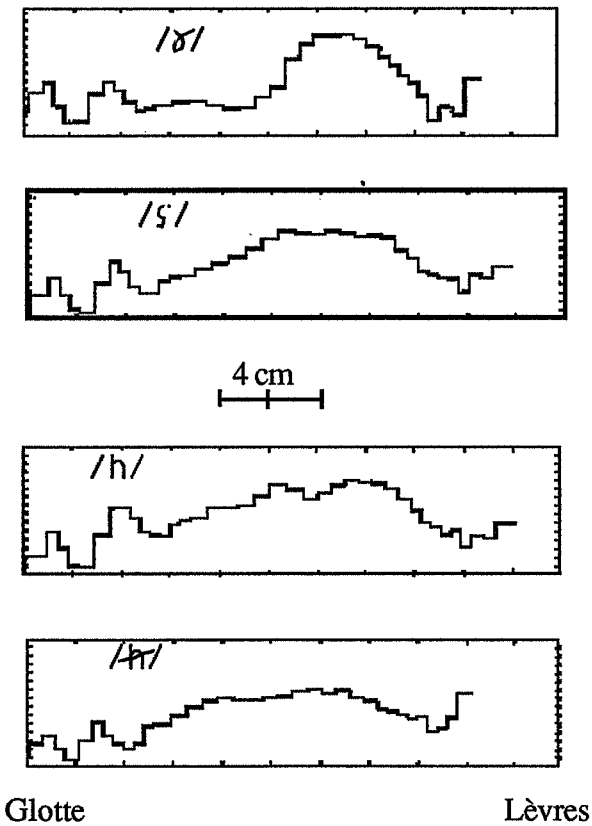


Fig.3. Coupes sagittales avec l'échelle de mesure La ligne médiane du cv est tracée en points tirés, en bas du graphe nous avons représenté le contour des lèvres pour les 5 fricatives d'arrière de l'arabe.



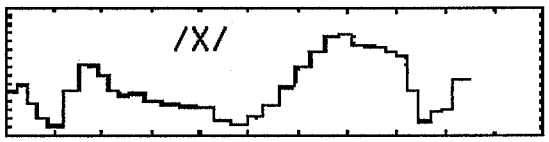


Fig.4.Les fonctions sagittales des 5 fricatives d'arrière de l'arabe

phase expérimentale : nous pensons que les valeurs de l'aire glottique : 0.05, 0.25, 0.35, 0.5 et 0.7cm^2 décrivent une plage suffisante pour la production des consonnes analysées.

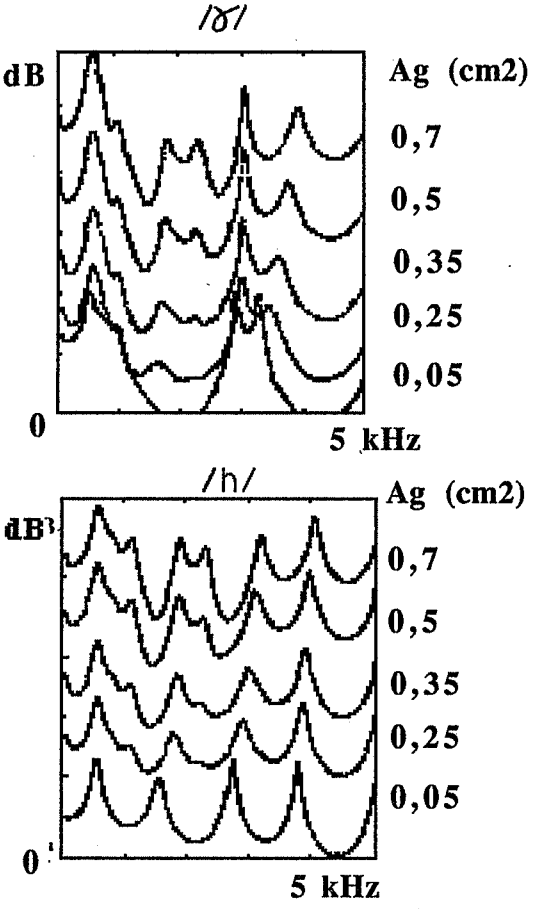
3.3 Les résultats

Il est bien visible sur les courbes de simulation que les effets subglottiques apparaissent d'autant mieux que la glotte est ouverte. Les pics spectraux correspondent de façon générale à ceux de la cellule de foster lorsque l'aire de la glotte se situe entre 0.25 et 0.35 cm^2, en fait cette ouverture varie aussi avec la forme des parties supraglottiques. C'est ainsi que dans le cas des pharyngales, l'aire glottique est plus voisine de 0.35cm^2, les fréquences obtenues sont 620, 990 et 1900Hz alors que dans le cas des uvulaires l'aire est plus proche de 0.25cm^2 et les fréquences obtenues sont

550, 1010 et 2050Hz. Dans ce dernier cas, nous avons relevé la présence systématique d'un pic que nous n'avons prévu ni pour les parties subglottiques ni pour les parties supraglottiques, sa fréquence est en moyenne égale à 1700Hz. Nous pouvons aussi remarquer sur nos courbes que la bande passante augmente lorsque la fréquence d'une partie suglottique est proche de celle d'une partie supraglottique. Par ailleurs, la présence des effets de couplage, entraine une différence dans le comportement des cavités supraglottiques : en effet les fréquences de résonance du conduit vocal ont augmenté.

Conclusion

Nous venons de développer une méthodologie qui a permis une évaluation des caractéristiques acoustiques du système phonatoire humain, c'est ainsi que nous avons pu obtenir de meilleures données sur les fonctions de transfert et les fonctions d'aire dans le cas des fricatives arrières. Nous avons pu aussi mettre en évidence les contributions acoustiques du système subglottique dans la production des consonnes fricatives arrières : en effet des pics spectraux supplémentaires sont observés clairement sur les fonction de transfert du conduit vocal à glotte à ouverte. Notre analyse systématique de ces effets sur chacune des configurations articulatoires a permis de mieux spécifier les relations acoustiques entre les parties subglottiques et les parties supraglottiques.



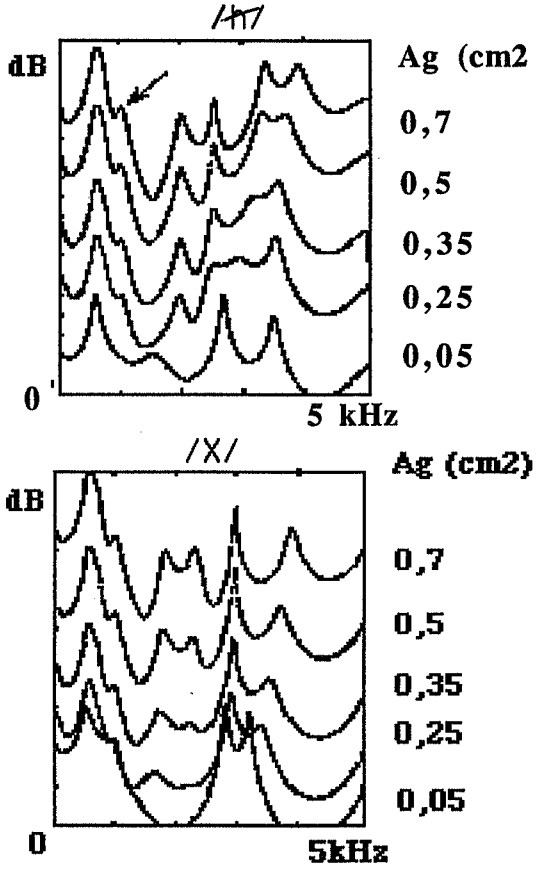


Fig.5. Fonctions de transfert simulées

Par ailleurs, l'étude a montré la possibilité de prendre en compte ces effets de couplage subglottique au niveau des simulations.

Références

Badin, P.(1989) Accoustics of voiceless fricatives:production theory and data,Speech transmission Laboratory,Quarterly Progress and Status Report No.3,33-55.

Bothorel, A., Simon, P., Wioland, F.& Zerling, J.P(1986) Cinéradiographie des voyelles consonnes du français, Travaux de l'Institut de Phonétique de Strasbourg.

Djéradi, A., Guérin, B., Badin, P., Perrier, P.(1991) Measurement of the acoustic transfer function of the vocal tract : a fast and accurate method, Journal of Phonetics 19, 387-395.

Fant, G. & Pauli,S. (1974) Spacial characteristics of vocal tract resonance mode, in Proceeding of the Second Speech Communication Seminar Stockholm

Mrayati, M. & Carre, R.(1976) Relation entre la forme du conduit vocal et les caractéristiques acoustiques des voyelles françaises, Phetica 33, 285-306.

Heinz, J.M. & Stevens, K.N.(1965) On the Relation between lateral cineradiographs, areara function and acoustic spectra of speech. In 5th International congress of Acoustic, Paper A44.

19es J.E.P. - Bruxelles, 19 au 22 mai 1992.

VOYELLES LONGUES ET VOYELLES BREVES EN ARABE STANDARD: ORGANISATION TEMPORELLE.

SALEM GHAZALI ET ABDELFATTAH BRAHAM

IRSIT ET UNIVERSITE DE TUNIS I, TUNIS.

Résumé

This paper is an attempt to investigate the temporal organization of short and long vowels in Modern Standard Arabic. Our main objectives are: a) to investigate the behavior of Arabic vowels with respect to the widely attested "closed syllable vowel shortening rule" (Maddieson 1985), b) to compare the duration of Arabic short and long vowels in different contexts and speech rates, and c) to test the hypothesis resulting from our feeling that speakers of Arabic tend to exaggerate the duration of long vowels and geminate consonants when pronouncing Standard Arabic in formal or educational settings to better signal quantity distinction.
The findings are discussed on the basis of two sets of data (105 words) read by three Tunisian university teachers of Arabic.

I. OBJECTIFS

Cette étude est une contribution aux nombreux travaux de recherche en théorie phonétique qui essayent de mieux comprendre l'organisation temporelle de la parole en général et de l'arabe en particulier. Les objectifs de ce travail sont de:
a. vérifier pour l'arabe standard l'hypothèse généralement admise pour plusieurs langues et selon laquelle une voyelle est plus brève en syllabe fermée qu'en syllabe ouverte [1] (Maddieson). Nous avons utilisé comme corpus des oppositions du type CaCCaCa-CaCaCa, CuCCiCa-CuCiCa et CaaCaC-CaaCCaC.

b. Comparer les durées des voyelles brèves et des voyelles longues dans différents contextes et étudier leur résistivité, et la résistivité des consonnes qui les suivent, à la compression en passant d'un débit normal de parole à un débit rapide.
c. Vérifier une hypothèse "intuitive" résultant de la constatation que, contrairement aux voyelles brèves, une voyelle longue en arabe standard peut avoir une marge de manoeuvre très importante. Les durées variables d'une voyelle longue pourraient être dues , entre autre, aux fonctions de cette langue dans une situation de diglossie.

II. METHODE EXPERIMENTALE

A.Les corpus

le premier corpus est constitué de quatre triplets:[kataba-kattaba-kaataba] [bada?a-badda?a-baada?a.], [ħallan-ħaalan-ħaallan] et [ʒaddan-ʒaadan- ʒaaddan].
Le deuxième corpus contient 90 mots et étudie les voyelles [a] [u] [aa] [uu] dans les schèmes "faʔala-faʔʔala-faaʔala et fuʔila-fuʔʔila-fuuʔila" où f = [n] et ʔ est une variable [b,d,t,t,k,q,f, s,s,ħ,z,ð,ð,l,n]. Dans les deux corpus, tous les mots choisis portent l'accent sur la syllabe expérimentale (la première syllabe) et sont inclus dans la phrase porteuse: [qaala ħamza ------- θalaaṯan] "Hamza a dit ----trois fois."
Il est peut être utile de rappeler à ce stade que l'arabe oppose aux voyelles brèves [a, u, i] les voyelles longues [aa, uu, ii]. Les voyelles

brèves peuvent apparaître en syllabe ouverte et fermée tout en étant plus centralisées et plus ouvertes en syllabe fermée [2]. Les voyelles longues [ii] et [uu] ne peuvent pas apparaître en syllabe fermée, seule [aa] est attestée dans ce contexte. Outre les oppositions de quantité vocalique, l'arabe utilise aussi les oppositions de quantité consonantique et chaque consonne peut être simple ou géminée à l'intérieur du mot.

B. Les sujets

Les sujets sont trois hommes agés de 38 à 40 ans (HR, SBR, et AB) tous locuteurs de l'arabe tunisien et enseignants d'arabe standard à l'université de Tunis I.
Le premier sujet (HR) a lu les deux corpus trois fois. Aucune instruction précise ne lui a été donnée concernant la façon dont il devait lire le corpus ou le but de la recherche. Ayant constaté que la fréquence fondamentale de sa voix est très élevée et variable et étant donné que le corpus devait servir pour d'autres investigations où F0 est une variable expérimentale, nous ne lui avons pas demander d'autres enregistrements.
Le deuxième sujet (SBR) a lu le premier corpus trois fois à une vitesse normale. Aucune instruction ne lui a été donnée avant ces premières lectures. Nous lui avons ensuite demandé de répéter cinq fois le premier et le deuxième corpus en débit normal et en débit rapide, mais en lui demandant cette fois-ci de parler naturellement et d'essayer de ne pas se comporter comme un "professeur d'arabe". En d'autres termes, nous lui avons dit, peut-être à tort, qu'il exagérait l'allongement de ses consonnes géminées et de ses voyelles longues pour marquer la distinction, comportement très fréquent dans les contextes pédagogiques.
Le troisième sujet (AB) a lu cinq fois le premier corpus dans les deux débits après avoir été averti des "abus" d'allongement de ses collègues professeurs d'arabe

C. Enregistrement et segmentation.

L'enregistrement a eu lieu dans une chambre insonorisée (artisanalement). Les données sonores ont été stockées initialement sur bande magnétique puis numérisées à 10KHz (12bits) à l'aide d'une carte d'acquisition (MacADIOS) et un logiciel d'analyse du signal de parole (MSL) fournis par GW Instruments et installés sur un MacintoshII.
Pour chaque mot, nous avons mesuré: a) la durée totale du mot, b) la durée de chaque segment, et c) la durée des cycles de closion CV, et de détente VC de la syllabe concernée [3]. Pour les voyelles, nous avons mesuré la phase vocalique seulement (de VOT à VVT).
Nous avons décidé d'écarter des résultats du présent travail les deux triplets du corpus 1 [kataba-kattaba-kaataba] et [ʒaddan-ʒaadan-ʒaaddan] car nous n'avons pas eu le temps de contrôler d'une façon précise le comportement des événements de transition telle que l'aspiration dans [kataba] ou le relâchement de [ʒ] dans [ʒaddan] d'une vitesse d'élocution à une autre.

Sujets et type de lecture	Débit normal	Débit rapide
H.R. lecture péd.	4,4 syl/sec	
S.B.R. lecture péd.	5,4 syl/sec	8,8 syl/sec
S.B.R. lecture ord.	6,8 syl/sec	
A.R. lecture Ord.	5,6 syl/sec	6,7 syl/sec

Tableau 1: Vitesse d'élocution par locuteur et par type de lecture, en débit normal et en débit rapide.

Enfin, pour calculer les vitesses d'élocution, seuls les mots objets de l'étude ont été pris en considération. Nous avons constaté que les sujets ralentissaient leur débit pour produire les mots cibles puis accéléraient pendant le reste de la phrase porteuse. La somme des durées de tous les mots cibles a donc été divisée par le

nombre total des syllabes qui les composent. Les résultats pour chaque sujet, exprimés en syllabes par seconde, sont détaillés dans le tableau 1. On peut noter qu'en débit normal, la vitesse d'élocution est moins rapide pendant les lectures considérées comme pédagogiques que pendant la lecture ordinaire. Ceci est en grande partie dû à l'allongement exagéré de VV et CC par rapport à V et C probablement pour bien signaler l'opposition de quantité.

II. RESULTATS ET DISCUSSION

1. [a] et [u] dans VC et VCC

Comme on peut le constater dans les tableaux 2, 3, 4 et 5 les voyelles brèves [a] et [u] s'abrègent en moyenne de 5ms dans le contexte VCC par rapport à VC (VCC/VC= 90% en moyenne). Ceci est vrai dans les deux corpus, pour les deux types de lecture (pédagogique et ordinaire) et quelle que soit la vitesse d'élocution, à l'exception du sujet HR. Celui-ci a présenté une différence plus importante (25ms) avant [d] (mais aucune différence avant [t], partie du corpus non utilisé dans ce travail).
L'examen détaillé des différences de durée des voyelles [a] et [u] en syllabes ouverte et fermée en fonction de la consonne postvocalique montre que le rapport VCC/VC est le plus élevé quand V=[a] et C une fricative voisée (97%) et le moins élevé quand C =[n] (82%). Avec [u] les variation (minimes) de durée ne semblent pas à première vue être associées à la nature de la consonne postvocalique.

	[a]	[aa]	Δ	[a]/[aa]
SBR	55ms	153ms	98ms	36%
HR	93ms	204ms	111ms	45%

Tableau 2 : Différence entre les durées de [a] et de [aa] en lecture "pédagogique".

2. [aa] dans VVC et VVCC

En débit normal, il y a deux tendances, selon le type de lecture. Dans une lecture pédagogique appliquée, la voyelle [aa] est réalisée plus longue en syllabe fermée VVCC qu'en syllabe ouverte VVC par les deux sujets.(aaCC/aaC=126% pour HR et 129% pour SBR). Dans une lecture ordinaire c'est l'inverse: la voyelle longue est plus brève en syllabe fermée aaCC qu'en syllabe ouverte aaC (aaCC/aaC=82% pour SBR et 93% pour AB). Cette tendance se maintient en débit rapide.(aaCC/aaC=88% pour les deux sujets).

3. [a] et [aa] en syllabe ouverte et en débit normal: Corpus 1

Nous comparons tout d'abord les valeurs des durées de [a] et [aa] devant l'occlusive voisée [d] chez les trois locuteurs dans le premier corpus. Chez les deux locuteurs qui ont effectué la lecture que nous avions qualifiée de "pédagogique" la différence de durée entre les deux voyelles est importante (tableau 2)
Ce rapport de durée entre brèves et longues est comparable à celui décrit par Port et al. (39%) pour l'arabe standard produit par des locuteurs égyptiens, irakiens et kowetiens et résumé dans le très intéressant travail de Jomâa sur l'organisation temporelle de l'arabe [4] .
Le même corpus prononcé par SBR durant une autre session d'enregistrement et après lui avoir demandé de lire d'une façon plus "ordinaire" a donné des rapports de durée différents. En effet, alors que la durée de la voyelle brève [a] est restée stable (55ms) celle de [aa] s'est considérablement comprimée (107ms) et le pourcentage de la brève par rapport à longue est passé à 51%.
Chez le troisième sujet (AB) qui, lui aussi, avait été prévenu des "abus" d'allongement dans un but pédagogique, on a obtenu presque le même rapport ([a]=66ms; [aa]=130ms; [a]/[aa]=50%).
Ces différentes conditions expérimentales suggèrent qu'une importante liberté de manoeuvre temporelle est réservée aux voyelles longues selon la situation à laquelle le locuteur fait face. Toutefois, le rapport V/VV n'a pas dépassé 51%.

4.[a]-[aa] , [u]-[uu] en syllabe ouverte et en débit normal: Corpus 2

Les durées des voyelles ouvertes brèves et longues, et des voyelles postérieures fermées sont comparées devant différents contextes consonantiques selon la prononciation de SBR effectuant une lecture "ordinaire" du deuxième corpus.

a. Les durées moyennes de [a] et [aa] (toutes consonnes confondues) sont comme suit: [a]= 61ms et [aa] =105ms ([a]/[aa] =58%).
b. Pour [u] et [uu] ces durées sont de 59ms et 96ms respectivement ([u]/[uu] =61%.
c. La durée de [aa] est légèrement supérieure (9ms) à celle de [uu], et la durée de [a] et celle de [u] sont presque identiques (2ms). Aucun test statistique n'a été appliqué pour déterminer si les différences sont significatives.
d. Le rapport [a]/[aa] est le plus élevé quand la consonne postvocalique est une fricative voisée (65%) et le moins élevé quand cette consonne est la liquide [l] (55%). Pour les autres consonnes ce rapport est autour de 58%.

lecture "pédagogique" et de 1,3 par rapport à sa lecture ordinaire. Pour AB, ce facteur est de 1,2 seulement par rapport à sa lecture ordinaire.
b. Dans le corpus 1, le rapport [a]/[aa] qui est chez SBR, en débit normal, de 36% pour une lecture pédagogique et de 51% pour la lecture ordinaire est passé en débit rapide à 71%. ([a] =47ms et [aa]=66ms). Il faut noter ici la forte compression de la voyelle longue: elle ne représente plus que 43% de sa valeur "pédagogique" et 60% de sa valeur "ordinaire", alors que la voyelle brève conserve 86% de sa valeur normale dans les mêmes conditions.
c. Chez AB, dont le facteur d'accélération est moins important, la voyelle longue [aa] représente en débit rapide, 80% de sa valeur en

	[a]				[aa]			
Sujets	VC	VCC	Δ	VCC/ VC	VVC	VVCC	Δ	VVCC/ VVC
H.R.	93ms	68ms	25ms	73%	120ms	152ms	32ms	126%
S.B.R. péd.	54ms	52ms	2ms	96%	147ms	191ms	44ms	129%
S.B.R. ord.	55ms	48ms	7ms	87%	109ms	89ms	20ms	82%
A.B. ord.	66ms	61ms	5ms	92%	136ms	127ms	9ms	93%

Tableau 3: Moyenne de la durée de [a] et de [aa] en syllabe ouverte et en syllabe fermée. (Débit normal - Corpus 1)

	[a]				[aa]			
Sujets	VC	VCC	Δ	VCC/ VC	VVC	VVCC	Δ	VVCC/ VVC
S.B.R.	47ms	42ms	5ms	89%	67ms	59ms	8ms	88%
A.B.	64ms	54ms	10ms	84%	107ms	95ms	12ms	88%

Tableau 4: Moyenne de la durée de [a] et de [aa] en syllabe ouverte et en syllabe fermée. (Débit rapide - Corpus 1)

5. [a], [aa] et [u], [uu] en syllabe ouverte et en débit rapide.
a. Les deux sujets qui ont lu les deux corpus avec différentes vitesses d'élocution sont SBR et AB. Le débit rapide de SBR représente un facteur d'accélération de 1,6 par rapport à sa

débit normal. Cette compression est négligeable pour la voyelle brève. (64ms contre 66ms).
d. Dans le corpus 2, les durées des voyelles ouvertes et fermées postérieures, longues et brèves sont comparées devant différents

contextes consonantiques selon la prononciation de SBR en débit rapide.

-a). Les durées moyennes de [a] et [aa] (toute consonnes confondues) sont comme suit: [a]=54ms, [aa]=78ms ([a]/[aa]=69%). Pour [u] et [uu] ces durées sont respectivement de 49ms et 74ms ([u]/[uu] =66%).

-b). La durée de [aa] est légèrement supérieure à celle de [uu] (4ms) et les durées de [a] et [u] sont légèrement différentes (5ms). Aucun test statistique n'a été appliqué pour déterminer si les différences sont significatives et si les taux de compression varient en fonction de la consonne postvocalique.

[uu] ne représentent plus, en débit rapide, que 74% et 77% de leur durée normale.

-c). Deux tests de reconnaissance (un test par débit) administrés au locuteur lui-même (SBR) et où les mots cibles ont été présentés dans un ordre aléatoire ont donné les résultats suivants: 100% de reconnaissance pour le débit normal quel que soit le type ou la longueur du segment. En débit rapide, ce taux de reconnaissance est de 93% pour [aa] et seulement de 53% pour [uu] qui se confond avec [u]. Ces taux seraient probablement plus bas si le sujet avait à reconnaître dans un même test une suite de mots prononcés à différentes vitesses d'élocutions (cf Jomâa, 1991 pour l'arabe tunisien.page 80).

	[a]				[u]			
Débit	VC	VCC	Δ	%	VC	VCC	Δ	%
normal	61ms	56ms	5ms	92%	58ms	52ms	6ms	89%
rapide	54ms	51ms	3ms	94%	49ms	42ms	7ms	85%

Tableau 5: Moyenne de la durée de [a] et de [u] en syllabe ouverte et en syllabe fermée, en débit normal et en débit rapide. (Corpus 2: tout contexte confondu. Sujet : S.B.R.)

corpus		a/ac		aa/aac		a/acc		u/uc		uu/uuc	
	Sujets / Débit	N	R	N	R	N	R	N	R	N	R
C2	S.B.R.	49%	49%	62%	59%	34%	37%	47%	46%	60%	57%
C1	S.B.R. péd.	47%	---	67%	---	28%	---	---	---	---	---
	S.B.R. ord.	48%	46%	63%	55%	31%	33%	---	---	---	---
	A.B.	56%	58%	72%	71%	32%	33%	---	---	---	---

Tableau 6: Taux de la durée vocalique par rapport au cycle de détente dans différent contexte et dans les deux débits. (C1 = corpus 1, contexte [d]; C2 = corpus2 tout contexte)

-c). Selon le débit, on constate que les voyelles brèves [a] et [u] ont respectivement conservé, en débit rapide, 88% et 84% de leur durée normale. Les voyelles longues [aa] et

6. Variations temporelles dans le cycle de détente.

Si on examine les tableaux 6 et 7 on constate que:

a. la phase vocalique des voyelles brèves reste relativement stable par rapport au cycle de détente quel que soit le débit. Le rapport V/VC tend à diminuer légèrement en accélérant la vitesse d'élocution (sauf pour AB où la tendance est vers la hausse). La compression intéresse les deux segments du cycle et semble uniforme, cependant la phase consonantique se
comprime plus que la phase vocalique lorsqu'il s'agit d'une consonne sourde. Quant au rapport V/VCC, il augmente légèrement en débit rapide.

Sujets et type de lecture	aa/aac		aa/accc	
	N	R	N	R
S.B.R. péd.	76%		60%	
S.B.R. ord.	69%	64%	53%	53%
A.B.	67%	65%	58%	53%

Tableau 7: Taux de la durée vocalique [aa] par rapport au cycle de détente [aal] et [aall].

b. pour les voyelles longues dans le cycle VVC la phase vocalique se comprime d'une façon plus importante en débit rapide. Cette compression par rapport au cycle de détente varie selon le sujet et en fonction de la nature de la consonne postvocalique. Toutefois, la phase vocalique constitue au moins 55% du cycle VVC en débit rapide, et ce même en syllabe fermée [aacc], où la phase vocalique occupe 53% du cycle VVCC.

La compression du segment long n'est pas spécifique à la voyelle. La consonne géminée manifeste aussi un taux de compression (26% en moyenne à l'exception de [l] et [n] où ce taux atteint 48%) supérieur à celui de la consonne simple (16% en moyenne, et 25% pour la liquide et la nasale).

REFERENCES

[1]. Maddieson, I. (1985). " Phonetic cues to syllabification" in *Phonetics Linguistics,Essays in Honor of Peter Ladefoged,* edited by V. Fromkin (Academic Press, New York) Pages 203-221.

[2]. Ghazali, S. (1979) " Du statut des voyelles en arabe". Analyse Théorie 2/3, pages 199-219

[3]. Abry, C., Benoit, C., Boe, J.P., et Sock, R. (1985). "Un choix d'événements pour l'organisation temporelle du signal de la parole". 14ème JEP du GCP de la SFA, pages 133-137.

[4] Jomâa. M., (1991) "Organisation temporelle acoustique et articulatoire de la quantité en arabe tunisien". Thèse de Doctorat , Université Stendhal, Grenoble.

L'APPORT DE LA CINÉMATIQUE DANS LA PERCEPTION VISUELLE DE L'ANTICIPATION ET DE LA RÉTENTION LABIALES

Marie-Agnès CATHIARD* & Mohamed Tahar LALLOUACHE**

***Laboratoire de Psychologie Expérimentale, U.A. CNRS N° 665, Université Pierre Mendès-France, BP 47X - 38040 Grenoble, France**
****Institut de la Communication Parlée, U.A. CNRS N° 368, INPG/ENSERG - Université Stendhal, BP 25X - 38040 Grenoble, France**

Résumé

Un phénomène de coarticulation, comme l'anticipation ou la rétention (persévération) labiale, qui est d'abord le reflet de stratégies à la production, semble pouvoir être perçu plus efficacement par la modalité visuelle que par la modalité auditive. Nous montrons de fait que l'anticipation naturelle du trait d'arrondissement en français lui permet d'être identifié de manière sûre (à 95%) jusqu'à 160 ms avant que le moindre signal acoustique ait été émis, et que la rétention de la position arrondie d'une voyelle pré-pausale peut être identifiée visuellement aussi bien encore 40 ms après le son. De plus, l'information récupérée visuellement n'est pas uniquement spatiale puisque nous avons pu mettre en évidence, en anticipation, un gain de 10 à 30 ms, apporté par le traitement du mouvement. En ce qui concerne la correspondance entre la cinématique de la production et sa perception, les «démarrages» de nos identifications en condition dynamique se sont révélés synchrones d'un événement cinématique remarquable, le pic d'accélération du geste de constriction/déconstriction qui contrôle l'aire à la sortie du conduit vocal.

1. INTRODUCTION

Anticipation et rétention (ou persévération) sont les deux versants temporels d'un phénomène proprement constitutif de la parole : la coarticulation. C'est un phénomène qui est sans aucun doute une commodité du point de vue du contrôle moteur, mais sans que l'on sache encore de manière décisive si c'est un avantage à la perception. Ainsi, en ce qui concerne la coarticulation labiale, en français – pour nous limiter à ce qui sera plus spécifiquement le sujet de notre étude – il a été démontré sur le plan articulatoire (de Benguérel & Cowan, 1974, à Abry & Lallouache, 1991) que le trait d'arrondissement d'une voyelle comme [y] peut être anticipé plus ou moins longtemps avant le début de son émission proprement voisée, et ce à travers plusieurs consonnes, dans des séquences de type [iC...Cy]. Mais, au niveau de l'information acoustique, l'efficacité perceptive de cette anticipation reste peu probante. Ainsi, dans une expérience subséquente de Benguérel & Adelman (1976), les sujets se sont révélés tout au plus capables d'identifier significativement (au-dessus du hasard), la voyelle à venir dans une telle séquence, seulement lorsqu'ils disposaient d'au moins la moitié du dernier segment consonantique. Plus généralement, il faut retenir que même un tenant, comme Whalen (1984), de l'efficacité perceptive des effets acoustiques de la coarticulation (pour faciliter la rapidité des décisions phonétiques) souligne que «nevertheless, the vowel information in stop bursts and frictions is quite weak. This is evident in our saying that these vowels can be identified at a "better than chance" level» (p. 49).

Le premier propos de notre communication sera de montrer, que l'efficacité perceptive *visuelle* de la coarticulation – en dépit du statut de cette modalité, longtemps considérée comme secondaire dans la parole (pour un débat récent exemplaire, cf. Massaro, 1989) – peut être autrement plus décisive que celle qui est obtenue généralement sur des signaux acoustiques. Nous avons choisi, pour cette démonstration, un cas limite : celui où le geste articulatoire d'arrondissement précède le signal acoustique de la voyelle de plusieurs dizaines de millisecondes, comme cela se produit dans la réalisation des pauses. Pour ce type de stimuli, qui sont par nature uniquement visuels sur la plus grande partie de leur décours (pour un premier résultat visuel en présence d'émission sonore, cf. Escudier et al., 1990), nous montrerons que les sujets sont capables d'identifier le trait d'arrondissement de manière fiable (à 95%), 60 à 160 ms avant le son, selon la durée de la pause. Ce bénéfice proprement visuel de la coarticulation n'avait, à notre connaissance, jamais été évalué, puisqu'on ne trouve dans la littérature qu'une brève mention d'une de ses expériences par McGurk (1981; expérience jamais publiée, comm. pers.), visant à mettre en évidence, par l'étude des temps de réaction, que le geste d'anticipation produit en début d'énonciation, peut être détecté sur des syllabes CV, avant que celles-ci ne soient auditivement perçues.

Notre second propos sera d'évaluer l'apport du *dynamique* ou, plus exactement, de la perception *cinématique* (on nous permettra d'utiliser indistinctement ces deux termes pour désigner le mouvement et/ou les forces impliquées). Notons qu'il y a actuellement dans les études sur la perception visuelle un net penchant en faveur du dynamique (Bonnet, 1984). Pour la parole, cette faveur se retrouve aussi bien dans les propositions de modélisation articulatoire (Browman & Goldstein, 1986; Saltzman & Munhall,

1989) que dans les expériences de perception auditive (Darwin, in Mattingly & Studdert-Kennedy, 1991). A partir des propositions de Strange & al. (1976), s'est même développé tout un débat sur l'importance du gain apporté par les informations coarticulatoires dynamiques des voyelles. Nous retiendrons de ce débat, qui est loin d'être clos, que c'est l'information proprement dynamique qui pourrait rendre compte d'un tel gain, que cette dynamique soit due à la coarticulation consonantique ou inhérente à la production de la voyelle (cette dernière proposition étant défendue par Andruski & Nearey, 1992). Parmi les métriques que Summerfield (1987) propose pour rendre compte de l'intégration audio-visuelle, une alternative qui lui semble prometteuse, serait de considérer que l'auditeur/labiolecteur pourrait identifier les paramètres pertinents des signaux optiques et acoustiques en termes de fonctions dynamiques plutôt qu'en termes d'états articulatoires successifs (ou cibles articulatoires). Pourtant dans les nombreuses expériences en perception visuelle, qui sont la plupart du temps réalisées avec des stimuli en mouvement (vidéo), les auteurs ne testent généralement pas l'apport du dynamique par rapport à la perception de stimuli statiques. En ce qui nous concerne, nous pourrons comparer les résultats en perception cinématique de la présente étude à ceux que nous avons précédemment obtenus en statique (Cathiard & al., 1991, 1992). Cette comparaison nous permettra ainsi d'estimer le *gain* apporté par la dimension temporelle par rapport à une information uniquement spatiale. On peut penser en effet que, par rapport à l'information sur une simple position, les sujets tireront davantage partie de la vue d'un geste dans son décours, pour en anticiper la suite, la vision du mouvement permettant d'accéder à des indices cinématiques tels que vitesse et accélération. Nous tenterons ainsi de mettre en correspondance nos résultats d'identification et certaines caractéristiques cinématiques que nous avons pu mesurer sur les lèvres.

2. METHODE

2.1. Corpus

Nous avons utilisé quatre transitions [V1#V2] avec (V1,V2 = [i] ou [y]; # étant la pause) qui étaient produites dans de petites phrases porteuses du type : «Tu dis: UHI ise?» et «T'as lu: IHU use?», où UHI, IHU, IHI et UHU sont, par convention, des prénoms d'indiens; «ise» et «use» étant en fonction de verbe. Les phrases [t y d i # i i i: z] et [t a l y # y y y: z] réalisent des transitions contrôle (que nous n'examinerons pas ci-dessous, faute de place). Avec dix répétitions, nous obtenons 40 phrases qui ont été enregistrées, en ordre aléatoire, par un locuteur, dans deux conditions (en deux prises de vues) : en marquant une pause courte [#]; puis, une pause plus longue [#:]. (Pour les conditions d'enregistrement vidéo et audio et le traitement des images, nous renvoyons à Cathiard & al., 1991, ainsi qu'à Lallouache, 1991, pour plus de détails)

2.2. Sélection des stimuli visuels

2.2.1. Mesures acoustiques

Après mesure de la durée des pauses intervocaliques, nous avons retenu deux réalisations, pour chacune des transitions, l'une représentative de la durée moyenne des pauses courtes (# = 160 ms) et l'autre représentative des pauses longues (#: = 460 ms).

2.2.2. Caractéristiques articulatoires

Parmi les mesures caractéristiques de la fente labiale (de face) et des protrusions des deux lèvres (de profil), nous avons retenu les évolutions temporelles de la protrusion de la lèvre supérieure (P1) et de l'aire (S), la première s'étant révélée généralement l'indicateur le plus direct du geste vocalique (ce qui n'est pas le cas pour tous les paramètres qui dépendent de la lèvre inférieure, laquelle subit les perturbations de la mandibule, Lallouache, 1991) et la seconde, la plus cruciale acoustiquement pour le maintien des effets de l'arrondissement en français (El Abed & Masmoudi, 1991). L'examen de P1 et S – ainsi que de leurs dérivées obtenues après lissage par fonctions splines cubiques – montre, pour la transition [i–>y] (Fig.1), une anticipation importante du geste d'arrondissement des lèvres. Après une première rampe plutôt lente, les mouvements de *protrusion* et de *constriction* [diminution de l'aire] démarrent – *d'après leurs pics d'accélération* (selon Perkell & Matthies, 1990) – respectivement, à 120 et 160 ms (pour la petite pause) et à 200 et 240 ms (pour la grande), avant le début acoustique du [y].

Par contraste, la transition [y–>i] (Fig.2), qui permet d'estimer la durée du *maintien* de l'arrondissement de [y] en position accentuée de fin de groupe prépausal (d'où le terme de *rétention* que nous avons finalement préféré à celui de *persévération* qui dénote plutôt une inertie), présente un décalage temporel des deux paramètres P1 et S, de même amplitude, mais *inverse*, puisque le mouvement de *rétraction* démarre 40 ms *en avance sur le relâchement de la constriction* (en petite pause, le pic d'accélération de P1 se situe 20 ms avant la fin acoustique du [y] et celui de S, 20 ms après; en grande pause, respectivement 20 et 60 ms après). Ce décalage reste, de plus, relativement constant sur la partie la plus rapide du geste (par comparaison, dans la transition [i–>y], P1 et S convergent plus vite, tout particulièrement dans la grande pause). Si nous estimons, par ailleurs, l'extension de la persévération de ces composantes P1 et S par leurs pics de décélération, nous obtenons, respectivement, des valeurs de 20 et 100 ms (pour la petite pause) et de 80 et 140 ms (pour la grande), à partir de la fin acoustique du [y].

En résumé, on a ainsi, quelle que soit la durée de la pause, dans [i–>y], une *anticipation* plus précoce de l'aire par rapport à la protrusion et dans [y–>i], une *rétention* plus longue de cette même aire. Mais, indépendamment de ce décalage inverse dans le *timing* des composantes du geste d'arrondissement, le phénomène majeur reste que, toutes choses égales par ailleurs, son anticipation est nettement plus longue que sa rétention.

2.3. Tests visuels

2.3.1. Test en condition statique

Nous avons testé, pour les transitions à petite pause, 13 images (dont 9 pendant le silence de la pause); et pour les transitions à grande pause, 28 images (dont

24 de pause); avec 3 images avant le début et une après la fin de chaque pause. Ceci nous donne un total de 164 stimuli, qui ont été présentés en ordre aléatoire, avec un décalage de 5 images à chaque changement de sujet. En outre, 4 images étaient présentées en début de test – sans *feedback* sur l'exactitude de la réponse – pour familiariser les sujets avec la tâche. Il s'agissait de vues de face présentées en noir et blanc, avec la partie supérieure du visage coupée à mi-hauteur des lunettes masquant les yeux, menton et cou étant visibles.

Chaque sujet, assis à environ un mètre de l'écran du moniteur, passait individuellement le test d'identification à choix forcé, le sujet devant décider pour chaque image si le locuteur prononçait un [i] ou un [y].

2.3.2. Test en condition dynamique

Les séquences d'images ont été constituées sur la base d'une fenêtre de 1200 ms, glissant par pas d'une image (soit 20 ms) : la dernière image de cette fenêtre était, tour à tour, une des images montrées dans le test statique et le début de la séquence se situait toujours en position préphonatoire. Afin d'avoir des tâches comparables dans les deux conditions statique et dynamique, nous avons réalisé, pour la condition dynamique, deux tests séparés de 82 stimuli chacun, l'un pour les séquences commençant par «tu dis...», l'autre pour les séquences commençant par «t'as lu...». (Le mélange des deux séquences dans un même test aurait en effet nécessité, de la part du sujet, l'identification de la séquence dans son ensemble et non pas seulement l'identification du seul son final de cette séquence.) Un fond gris uniforme, de luminance équivalente à la luminance moyenne d'une image, a été généré et présenté en permanence à l'écran avant et après l'apparition des séquences, ceci afin que l'œil puisse s'adapter à la luminance des stimuli. Un bip sonore était programmé 3 secondes avant l'apparition de chaque séquence afin de recapter l'attention du sujet (la tâche étant nettement plus attentionnelle dans cette condition dynamique, en raison de la brièveté du signal). Le moniteur utilisé pour la visualisation des séquences était à rémanence faible, inférieure à 20 ms, pour pouvoir tester des identifications en dynamique à la trame près.

Les deux tests correspondant à chaque phrase étaient présentés au sujet au cours d'une même session : un contre-balancement de l'ordre de passation des tests a été réalisé à chaque nouveau sujet, ainsi qu'un décalage de 5 images par rapport à l'ordre aléatoire initial des tests. Le sujet devait identifier le son final de chaque séquence : pour les séquences «tu dis...», [y] dans [t y d i # y...] et [i] dans [t y d i # i...] ou [i] par défaut, s'il n'avait vu que [t y d i...]; pour les séquences «t'as lu...», [i] dans [t a l y # i...] et [y] dans [t a l y # y...] ou [y] par défaut, s'il n'avait vu que [t a l y...]. Six séquences de familiarisation (sans *feedback* sur l'exactitude de la réponse) étaient programmées en début de chaque test.

2.4. Sujets

50 étudiants (25 pour chaque test), de langue maternelle française, ont participé aux tests d'identification. Aucun sujet n'avait de compétence particulière en lecture labiale, ni n'avait reçu d'enseignement en phonétique. Des tests d'acuité visuelle et auditive ont été effectués avant passation, tous les sujets retenus ayant une vue normale ou parfaitement corrigée ainsi qu'une audition correcte.

3. RESULTATS

Nous présenterons nos résultats sous forme de courbes d'identification, pour lesquelles nous déterminerons les frontières à 50% et effectuerons une comparaison en condition statique et en condition dynamique. Enfin, nous confronterons nos résultats perceptifs à nos données articulatoires.

3.1. Statique *vs.* dynamique

Nous avons calculé, pour les 25 sujets de chaque condition, le pourcentage d'identification [y] obtenu pour chaque image statique et pour chaque image finale des séquences dynamiques. Les courbes d'identification, dressées pour chaque transition [i–>y] et [y–>i], présentent une forme classique en sigmoïde.

3.1.1. Transition [i–>y] (Fig.1)

La perception visuelle peut être efficace plusieurs dizaines de millisecondes avant même que le moindre signal acoustique ait été émis, puisque les sujets identifient correctement l'information segmentale d'arrondissement (95% de réponses [y]), en conditions statique et dynamique, respectivement 40 et 60 ms avant le début acoustique du [y], dans la petite pause, et 120 et 160 ms, dans la grande.

Pour déterminer la *frontière perceptive visuelle* sur nos fonctions d'identification, nous avons utilisé l'analyse Probit de Finney (1971) qui permet : (i) d'estimer pour chaque courbe, la position temporelle de sa frontière à 50% (sa moyenne) et sa pente, en utilisant toutes les données comprises entre les régions asymptotiques; (ii) de comparer les frontières des courbes, par un test de type Student.

Les ajustements effectués montrent que la frontière se situe : pour la petite pause, à 90 ms avant le début acoustique du [y], en condition statique, et à 100 ms, en condition dynamique; et pour la grande pause, à 180 ms et à près de 210 ms, respectivement.

Pour chaque condition, il y a une différence significative ($p<.01$) des frontières obtenues en petite et grande pauses : globalement, *quand la pause triple, l'anticipation visuelle double;* la position temporelle de la frontière dépend donc de l'extension en durée de l'anticipation articulatoire.

Pour chaque valeur de pause, la comparaison des frontières, entre les conditions statique et dynamique, révèle une différence de 10 ms (significative à $p<.05$) pour la petite pause et de 30 ms (significative à $p<.01$) pour la grande pause. *L'identification du geste d'arrondissement est donc plus précoce en condition dynamique qu'en condition statique.*

Nous pouvons aussi remarquer que les courbes obtenues en condition dynamique sont moins bruitées, ce qui explique sans doute leurs pentes un peu plus raides ($b = 0.77$ pour la petite pause et $b = 0,91$ pour la grande pause) par rapport aux pentes des courbes obtenues en condition statique ($b = 0.62$ et 0.65, respectivement). Cette différence de «lissé» ne conduit d'ailleurs pas au rejet de l'hypothèse de parallélisme entre les courbes : la *largeur ou précision temporelle* de

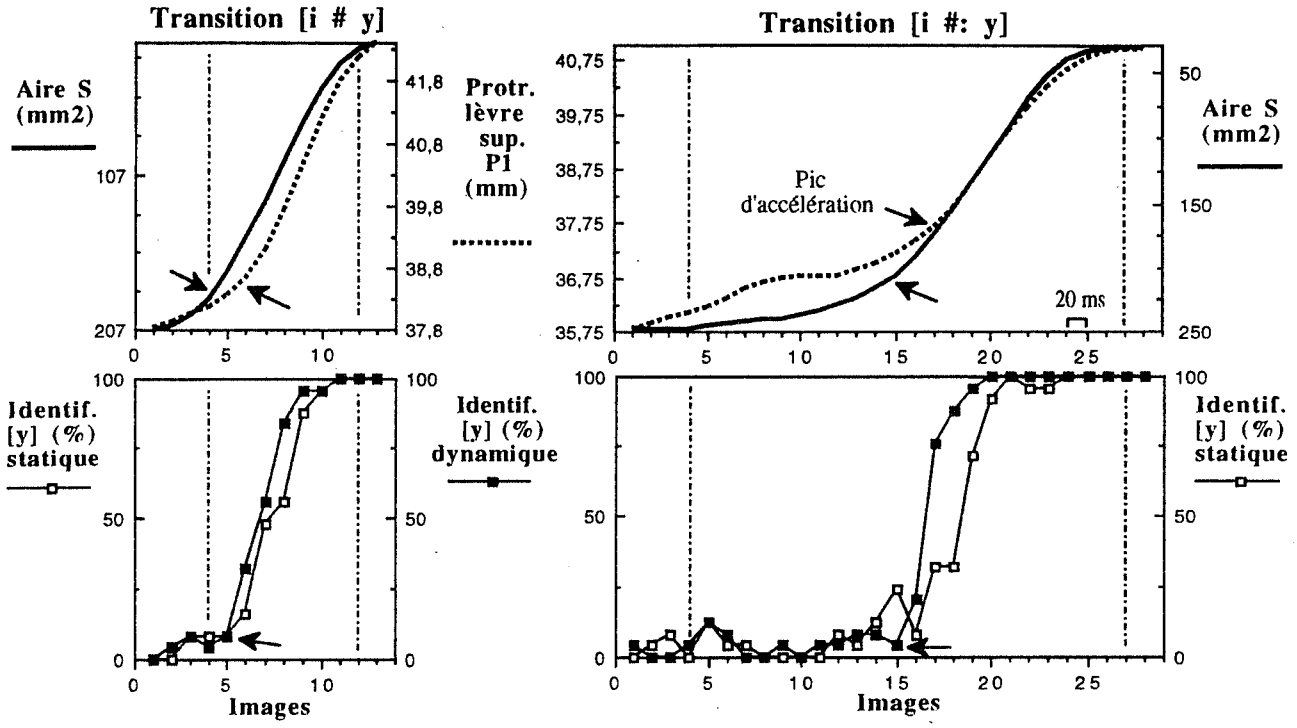


Figure 1. : Transitions [i -> y] avec petite (#) et grande pause (#:)
– En haut : Évolution de la protrusion de la lèvre supérieure (P1) et de l'aire (S), après lissage par fonctions splines cubiques (les flèches indiquent les pics d'accélération).
– En bas : Fonctions d'identification [y] correspondantes, pour 25 sujets français, en conditions dynamique et statique (sur les courbes dynamiques, les flèches indiquent les «démarrages» repérés par les pics de dérivées seconde). La verticale point-tiret de gauche indique la fin acoustique du [i]; celle de droite, le début du [y].

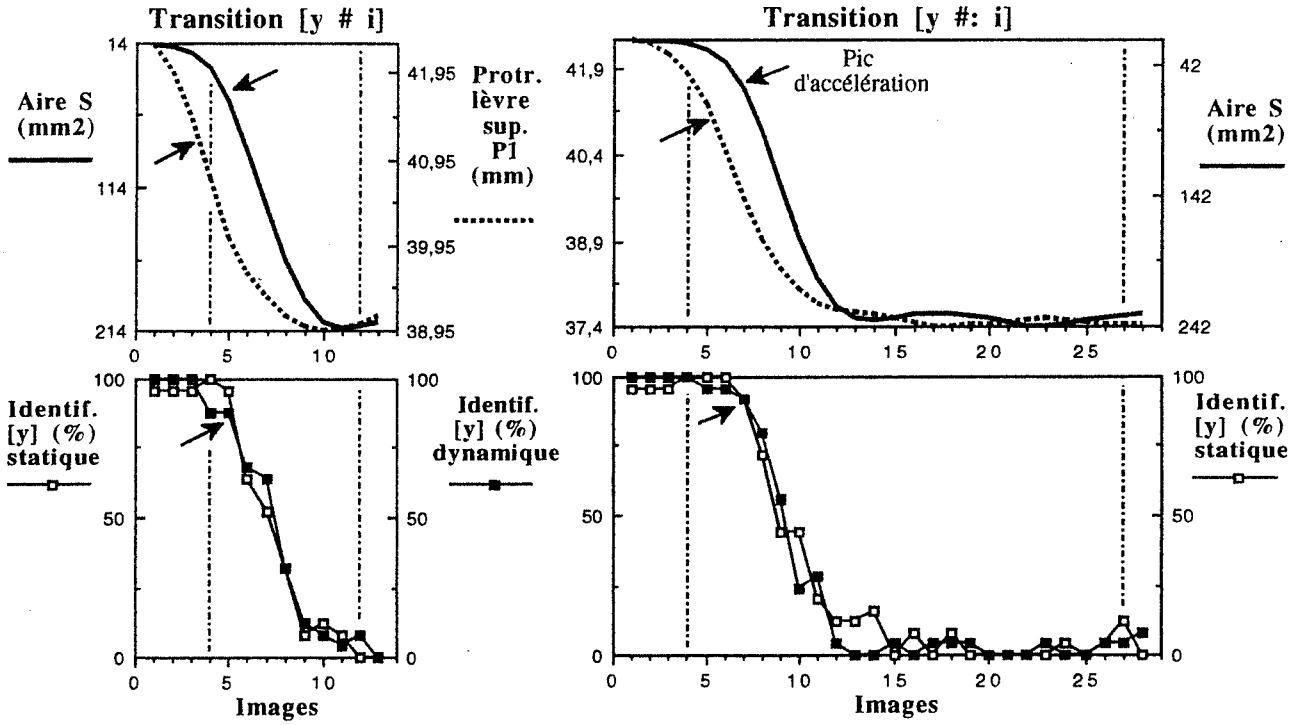


Figure 2. : Transitions [y -> i] avec petite (#) et grande pause (#:)
– En haut : Évolution de la protrusion de la lèvre supérieure (P1) et de l'aire (S), après lissage par fonctions splines cubiques (les flèches indiquent les pics d'accélération).
– En bas : Fonctions d'identification [y] correspondantes, pour 25 sujets français, en conditions dynamique et statique (sur les courbes dynamiques, les flèches indiquent les «démarrages» repérés par les pics de dérivées seconde). La verticale point-tiret de gauche indique la fin acoustique du [y]; celle de droite, le début du [i].

nos frontières reste relativement stable, puisqu'aussi bien en dynamique qu'en statique, dans la petite comme dans la grande pause, *80 ms* environ suffisent pour basculer de [i] à [y].

3.1.2. Transition [y->i] (Fig. 2)

La perception visuelle de l'arrondissement, dans la transition de la voyelle [y] vers la voyelle [i], à travers la pause, peut persévérer plus ou moins après la fin de l'émission acoustique de la voyelle : à 95 % de réponses correctes, les sujets peuvent continuer à identifier [y] jusqu'à 40 ms après la fin acoustique de la voyelle (pour la petite pause, -20 ms en condition dynamique, 20 ms en statique, et 40 ms dans la grande pause pour les deux conditions).

La frontière visuelle des courbes estimée à 50% par Probit se situe 60 ms après la fin acoustique du [y] dans la petite pause, et près de 110 ms après dans la grande pause, en conditions statique et dynamique.

La différence, entre les deux durées de pause, pour les deux conditions de test, s'avère significative ($p<.01$). Là aussi, comme pour l'anticipation – et même si les durées de rétention sont nettement plus faibles – on peut dire que, globalement, *lorsque la pause triple, la rétention visuelle double.*

Pour chaque valeur de pause, la comparaison des frontières, entre les conditions statique et dynamique, ne révèle aucune différence significative. Quant à la précision temporelle des frontières, elle est identique à celle observée dans la transition [i–>y].

3.2. Identifications et cinématique des lèvres

Quelle est la correspondance entre ces courbes d'identification et les fonctions temporelles articulatoires, autrement dit entre la perception et la production de nos stimuli?

Pour la transition [i–>y], l'observation comparée des courbes d'identification d'une part, et de l'évolution de la protrusion (P1) et de la constriction (S) des lèvres d'autre part, indique que les identifications ne suivent pas le décours de ces paramètres. La première phase lente du geste n'est en effet nullement prise en compte par nos sujets. Quant à la seconde phase plus rapide, qui atteindra son maximum seulement 20 ms avant le début acoustique du [y], elle est largement accentuée puisque nos sujets peuvent atteindre 95% d'identification correcte, 160 ms en avance sur le son. Nous avons comparé événements cinématiques (pics de vitesse et d'accélération de P1 et S) et singularités des fonctions d'identification (leurs pics de dérivées première et seconde), en dynamique. Cette comparaison indique qu'il y a correspondance à une image près, pour la petite pause, entre le pic de dérivée seconde de l'identification et le pic d'accélération de l'aire (le pic sur l'identification, qui se situe 140 ms avant le début acoustique du [y], étant en retard de 20 ms sur celui de l'aire) et synchronie pour la grande pause (les deux pics se situant à 240 ms du début acoustique du [y]).

Cette coïncidence d'événements cinématiques se retrouve aussi pour la transition [y–>i], puisque pic de dérivée seconde de l'identification et pic d'accélération du geste de relâchement de la constriction sont parfaitement synchrones, en petite pause (ces pics se situant 20 ms après la fin acoustique du [y]), comme en grande pause (les pics se situant 60 ms après la fin acoustique du [y]).

Cette coïncidence généralisée sur l'aire est d'autant plus remarquable que dans les transitions [y->i], l'avance du geste de rétraction sur le relâchement de la constriction – une avance qui se maintient, rappelons-le, dans la partie la plus rapide de la transition vocalique – n'a pas davantage été perçue en condition dynamique qu'elle ne l'avait été en condition statique. Il se pourrait, bien entendu, que le geste de protrusion/rétraction soit globalement difficile à percevoir sur une présentation *de face* de nos stimuli, alors que les variations d'aire aux lèvres seraient aisément repérables. Une autre hypothèse serait que nos sujets ne tiendraient pas compte des variations de la protrusion mais uniquement de celles de l'aire aux lèvres, cette variation se révélant la plus directement efficace acoustiquement pour passer de [y] à [i] en français (El Abed & Masmoudi, 1991). Une expérience en cours, avec une présentation de profil des mêmes stimuli, devrait nous permettre de tester la première hypothèse, la seconde devant l'être avec des stimuli synthétiques.

4. CONCLUSIONS

Dans cette étude sur la perception visuelle de l'anticipation et de la rétention du trait d'arrondissement en français, nous avons utilisé le contrôle de la durée des pauses acoustiques. Dans la perception de la parole, ces pauses – qui ont une fonction prosodique évidente (tout particulièrement en parole dite «claire», Cutler & Butterfield, 1990) – nous offrent un cas naturel où l'information *segmentale* à venir ne peut pas être récupérée pendant le silence autrement que par la modalité visuelle, cette dernière pouvant ainsi tirer parti d'un flux optique ininterrompu.

L'efficacité perceptive de cette modalité est démontrée, en premier lieu, par le fait que *l'anticipation naturelle du trait d'arrondissement lui permet d'être identifié de manière sûre (à 95%), jusqu'à 160 ms avant que le moindre signal acoustique ait été émis.* Mais aussi par le fait que *la rétention de la position arrondie d'une voyelle pré-pausale peut être visuellement perçue encore 40 ms après le son.* Ces résultats peuvent s'inscrire plus généralement dans une démonstration de la supériorité des informations visuelles sur les informations auditives en ce qui concerne la coarticulation, puisque ces dernières n'atteignent jamais des scores aussi élevés (Whalen, 1984), ni n'offrent en particulier une telle avance (Benguérel & Adelman, 1976). Ce qui est vrai *a fortiori* pour l'anticipation dans nos pauses silencieuses, devrait l'être dans le cas de suites [iC...Cy]. Notons que, dans une langue comme le français, qui présente le même visème pour [y] et [u] (Tseva, 1990), cet avantage pourrait être néanmoins exploité pour un accès lexical, sinon par segments, du moins par traits (Lahiri & Marslen-Wilson, 1991).

L'information récupérée visuellement n'est pas une information uniquement spatiale puisque nous avons mis en évidence, *en anticipation, un gain de 10 à 30 ms, apporté par le traitement du mouvement.* Ce gain, bien que significatif, correspond en moyenne à une

avance d'une image et n'atteint pas, même dans le meilleur cas, deux images. Nous nous trouvons donc devant un effet qui est loin d'être massif et dont la robustesse reste bien évidemment à tester par d'autres expériences. Il faut noter à ce propos que la faveur dans laquelle sont tenues les informations dynamiques en parole (Strange & al. ,1976, Darwin, in Mattingly & Studdert-Kennedy, 1991, Andruski & Nearey, 1992), et bien entendu en perception visuelle, ne requiert pas forcément la présentation de stimuli dynamiques, puisque Freyd (1983) a pu prouver, par une série d'expériences, que le mouvement pouvait être perçu à partir de stimuli *statiques*.

Enfin, en ce qui concerne la correspondance entre la cinématique de la production et sa perception, on a pu montrer que les «démarrages» de nos identifications en condition dynamique sont synchrones d'un événement cinématique remarquable, le pic d'accélération du geste de constriction/déconstriction des lèvres, qui contrôle l'aire à la sortie du conduit vocal, un paramètre dont l'importance acoustique n'est plus à démontrer. Ce dernier résultat nous invite à explorer le traitement de l'accélération visuelle pour ce type de mouvements dont les buts acoustiques sont linguistiquement cruciaux en français.

Plus généralement, nous pouvons nous demander pourquoi ces phénomènes de coarticulation, qui ne sont après tout que le reflet de stratégies à la production, peuvent être utilisés au mieux dans la perception (pour un récent débat sur la théorie motrice de la parole, cf. Summerfield in Mattingly & Studdert-Kennedy, 1991).

Remerciements à Willy Serniclaes et à Jean-Luc Schwartz pour nous avoir guidés dans l'analyse Probit; à Claude Bonnet, pour nous avoir conseillés sur les caractéristiques optiques à adopter pour notre expérience de *gating*; enfin, à Jacques Droulez pour avoir attiré notre attention sur les travaux de Freyd.

REFERENCES

Abry, C. & Lallouache, M.T. (1991). Audibility and stability of articulatory movements : deciphering two experiments on anticipatory rounding in French, in *Proceedings of the XIIth International Congress of Phonetic Sciences*, 19-24 Août 1991, Aix-en-Provence, France, 1, 220-225.

Andruski, J.E. & Nearey, T.M. (1992). On the sufficiency of compound target specification of isolated vowels and vowels in /bVb/ syllables, *JASA, 91(1)*, 390-410.

Benguérel, A.P. & Adelman, S. (1976). Perception of coarticulated lip rounding, *Phonetica, 33*, 113-126.

Benguérel, A.P. & Cowan, H.A. (1974). Coarticulation of upper lip protrusion in French, *Phonetica, 30*, 41-55.

Bonnet, C. (1984). *Psychophysique de la perception visuelle du mouvement*, Thèse de Doctorat d'Etat, Université Pierre et Marie Curie, Paris 6.

Browman, C.P. & Goldstein, L. (1986). Towards an articulatory phonology, *Phonology Yearbook, 3*, 219-252.

Cathiard, M.-A., Tiberghien, G., Cirot-Tseva, A., Lallouache, M.-T. & Escudier, P. (1991). Visual perception of anticipatory rounding during acoustic pauses : a cross-language study, in *Proceedings of the XIIth International Congress of Phonetic Sciences*, 19-24 Août 1991, Aix-en-Provence, France, *4*, 50-53.

Cathiard, M.-A., Cirot-Tseva, A. & Lallouache, M.-T. (1992). Identification visuelle des gestes de protrusion et de rétraction des lèvres au cours des pauses acoustiques, *Deuxième Congrès Français d'Acoustique*, 14-17 Avril 1992, Arcachon, France.

Cutler, A. & Butterfield, S. (1990). Durational cues to word boundaries in clear speech, *Speech Communication, 9*, n°5/6, 485-495.

El Abed, R. & Masmoudi, I. (1991). *Modélisation articulatori-acoustique de la transition [y–>i] en français.* DEA Sciences du Langage, Université Stendhal, Grenoble.

Escudier, P., Benoit, C. & Lallouache, M.-T. (1990). Identification visuelle de stimuli associés à l'opposition /i/–/y/ : étude statique, Premier Congrès Français d'Acoustique, *Colloque de Physique, suppl. au n°2, Tome 51*, C2-541-544.

Freyd, J.J. (1983). The mental representation of movement when static stimuli are viewed, *Perception & Psychophysics, 33(6)*, 575-581.

Finney, D.J. (1971). *Probit analysis*, Cambridge University Press.

Lahiri, A. & Marlsen-Wilson, W. (1991). The mental representation of lexical form: a phonological approach to the recognition lexicon, *Cognition, 38*, 245-294.

Lallouache, M.-T.(1991). *"Un poste 'visage-parole' couleur. Acquisition et traitement automatique des contours des lèvres"*, Thèse de l'INP, Grenoble.

McGurk, H. (1981). Listening with eye and ear (paper discussion), in T. Myers, J. Laver & J. Anderson (Eds.), *The cognitive representation of speech*, Amsterdam, North-Holland, 336-338.

Massaro, D.W. (1989). Multiple book review of Speech perception by ear and eye : a paradigm for psychological inquiry, *Behavioral and Brain Sciences, 12*, 741-794.

Mattingly, I.G. & Studdert-Kennedy, M. (Eds.) (1991), *Modularity and the motor theory of speech perception*, Lawrence Erlbaum Associates.

Perkell, J.S. & Matthies, M.L. (1990). Timing of upper lip protrusion gestures for the vowel /u/, *JASA, Suppl. 1, 87*, S123.

Saltzman, E.L. & Munhall, K.G. (1989). A dynamical approach to gestural patterning in speech production, *Ecological Psychology, 1*, 333-382.

Strange, W., Verbrugge, R.R., Shankweiler, D.P. & Edman, T.R. (1976). Consonant environment specifies vowel identity, *JASA, 60(1)*, 213-224.

Summerfield, Q. (1987). Some preliminaries to a comprehensive account of audio-visual speech perception, in B. Dodd & R. Campbell (Eds.), *Hearing by eye : the psychology of lipreading*, L E A. , 3-51.

Tseva, A. (1990). Les visèmes vocaliques du français : une analyse multidimensionnelle, *Bulletin d'Audiophonologie, 6*, 3/4, Annales scientifiques de l'Université de Franche-Comté, 381-411.

Whalen, D.H. (1984). Subcategorical phonetic mismatches slow phonetic judgements, *Perception & Psychophysics, 35*, 49-64.

19es J.E.P. - Bruxelles, 19 au 22 mai 1992.

PHONÉTIQUE, PHONOLOGIE ET ART VERBAL

MARC DOMINICY

UNIVERSITÉ LIBRE DE BRUXELLES

Les traités de versification s'ouvrent souvent par une déclaration liminaire qui proclame que les phénomènes métriques "trouvent leur source" dans certaines propriétés de la langue. Cependant, peu d'auteurs se risquent à préciser la nature de cette relation. Et à lire les citations rassemblées par Allen (1973: 12-13), on ne peut s'empêcher de croire que, derrière cette évidence unanimement répétée, se dissimulent des problèmes parfois très troublants. Certes, on ne saurait guère reprocher à T.S. Eliot de recourir à un aphorisme plus brillant que démonstratif quand il écrit que:

> "The music of poetry must be a music latent in the common speech of its time"

Par contre, il est davantage gênant de lire à peu près la même chose sous la plume d'un phonéticien comme Abercrombie:

> "The rhythm of everyday speech is the foundation of verse, in most languages"

Car de ce genre d'affirmation, un esprit non prévenu pourrait aisément conclure que les conventions métriques se laissent ramener à des régularités phonétiques immédiatement corrélées à la récitation spontanée du vers. Or, on sait depuis toujours qu'il convient de faire une distinction rigoureuse entre le vers et son "exécution", ne serait-ce que pour caractériser les "modèles d'exécution" (les normes de récitation du vers) qui ont pu être privilégiés à tel ou tel moment de l'histoire (Jakobson 1960). On sait également que certains modèles d'exécution, comme la "restitution métrique" de l'alexandrin français, présupposent des réalisations phonétiques (des diérèses ou des synérèses, par exemple) qui sont étrangères à toute prononciation synchroniquement attestée.

De manière plus générale, les métriciens contemporains (voir, par exemple, Cornulier 1982) ont dénoncé, à juste titre, l'illusion positiviste qui consisterait à réduire la perception des parallélismes métriques à un simple repérage d'indices phonétiques nécessairement présents dans une lecture "correcte" ou "légitimement expressive" du vers. Mais si les recherches métriques les plus récentes ont souvent échappé au réductionnisme phonétique, elles n'offrent cependant aucune réponse clairement articulée à la question que j'évoquais ci-dessus. En d'autres termes, nous n'avons encore aucune idée très nette sur la manière dont les conventions métriques "s'enracinent" dans le système linguistique, et sur les contraintes qui découlent de cette dépendance.

Je ne prétends pas fournir ici une réponse définitive. Ce que j'essaierai de faire, plutôt, c'est de critiquer une conjecture qui se recommande par sa simplicité et par son haut degré de falsifiabilité, et que j'appellerai, pour faire bref, l'"hypothèse de l'équivalence". Cette conjecture a été avancée par Verluyten (1989: 32) sous la forme qui suit:

(1) Toute catégorie métrique a son équivalent dans une catégorie prosodique de la langue.

Pour bien saisir la portée de pareille affirmation, il faut immédiatement la replacer dans son contexte.

Ce que soutient Verluyten n'exige pas, tout d'abord, que l'échelle de toutes les valeurs possibles d'une catégorie métrique donnée soit isomorphe à l'échelle de toutes les valeurs possibles de la catégorie prosodique correspondante. Ainsi, dans une métrique acccentuelle (comme celle de l'anglais), la catégorie métrique "accent" ne retient que deux valeurs ("accentué" vs. "inaccentué"), alors que la catégorie prosodique équivalente (l'accent de mot, primaire ou secondaire, principal ou subsidiaire) connaît une échelle à plus de deux valeurs. C'est la raison pour laquelle le mot *happily*, par exemple, pourra être considéré comme métriquement binaire (avec un accent sur la première et la troisième syllabes) ou métriquement ternaire (avec un accent sur la première syllabe seulement), d'après le

traitement réservé à l'accent subsidiaire porté par la troisième syllabe. De même, dans la versification chinoise classique, la catégorie métrique du "ton" ne retient que les deux valeurs "étale" (ou "plat") et "défléchi" (ou "modulé"), alors que l'échelle associée à la catégorie prosodique équivalente renfermait un nombre bien plus élevé de valeurs (cf. Lotz 1960: 141). Je m'interrogerai plus loin sur les raisons qui déterminent ce mécanisme de binarisation métrique, dont Jakobson (1960) a montré le caractère systématique.

Par contre, l'hypothèse de Verluyten nous interdit de poser une catégorie métrique à laquelle nous ne pourrions trouver aucun équivalent prosodique. Il s'agit là d'une exigence sans doute excessive, qui nous contraint d'emblée à de périlleuses approximations. Soit, par exemple, des catégories métriques aussi traditionnelles que celles de pied, d'hémistiche, de vers ou de strophe. On a souvent soutenu — et Verluyten reprend ce genre de thèse — qu'elles trouvaient, dans la langue, des équivalents prosodiques tels que (respectivement) le mot, le constituant, la phrase et l'unité discursive. Cependant, l'expérience montre rapidement que, dans pareils cas, les textes poétiques accumulent les discordances entre les catégories métriques et leurs prétendus équivalents prosodiques. En d'autres termes, il est banal qu'à un pied ne corresponde pas un mot, ou qu'à un vers ne corresponde pas une phrase, alors que les discordances entre, par exemple, l'une des deux valeurs de l'accent métrique et l'une des valeurs de l'accent de mot obéissent à des contraintes beaucoup plus sévères.

A bien y réfléchir, cette dernière conclusion ne doit pas trop nous surprendre. Si des catégories comme le mot, le constituant, la phrase, voire l'unité discursive peuvent, à la rigueur, passer pour "prosodiques", il leur manque, en tous cas, une propriété que possèdent l'accent aussi bien que le ton, à savoir la dimensionnalité (l'existence d'au moins deux valeurs figurant sur l'échelle associée à la catégorie). C'est cette dimensionnalité, nous l'avons vu, qui permet, le cas échéant, de construire la catégorie métrique équivalente par binarisation. Quant aux causes profondes de la binarisation elle-même, elle me paraissent résider dans le fait qu'un texte poétique combine deux niveaux d'organisation bien distincts, de sorte qu'un sujet acculturé à la création, à la réception ou à la restitution de ce type d'objet doit gérer des problèmes d'optimalisation parfois complexes. Il est clair qu'une stratégie qui rabat une catégorie multidimensionnelle (de degré supérieur à 2) sur une catégorie binaire laisse plus de latitude à l'expression qu'une stratégie ayant à relier deux catégories multidimensionnelles mutuellement irréductibles.

Il semble donc judicieux de restreindre l'hypothèse de Verluyten aux catégories prosodiques dimensionnelles, qui sont susceptibles de se prêter à la binarisation. Or, ce que je voudrais montrer ici, c'est que même une version aussi timide de la conjecture (1) se heurte à de véritables contre-exemples.

Afin de justifier (1), Verluyten reprend l'argument bien connu, selon lequel l'échec de toutes les tentatives qui ont pu être menées pour implanter une métrique quantitative en français s'explique par l'absence, dans cette langue, de la catégorie prosodique de "poids", qui répartit les syllabes en au moins deux classes complémentaires (syllabes "lourdes" et syllabes "légères"). Par contraposition, ce raisonnement entraîne que la catégorie prosodique en question devait être présente, synchroniquement, dans les systèmes du latin classique et du grec ancien. Nous ne rencontrons guère de problèmes pour le latin (où, pourtant, la métrique quantitative a été largement ou totalement importée du grec), puisque le poids syllabique y détermine la place de l'accent au sein des mots contenant plus de deux syllabes (cf. Allen 1973, Zirin 1970). Par contre, les données concernant le grec ancien soulèvent de très nombreuses difficultés, dans la mesure où, selon toute probabilité, aucun phénomène synchronique ne faisait intervenir la catégorie prosodique de poids.

On sait, tout d'abord, que l'accent (libre) du grec ancien obéissait à des contraintes de placement qui ne connaissaient pratiquement pas d'exceptions. A la suite de Allen (1973: 230-239), je représenterai l'accent grec par une "contonation" (une mélodie) Haut-Bas, que je noterai "HB". Il est alors possible de remplacer les "règles de limitation" figurant dans les traités traditionnels par trois principes fort simples:

(2) H et B s'attachent à la même syllabe ou à deux syllabes immédiatement contiguës;
(3) H ne peut s'attacher au-delà de la syllabe antépénultième;
(4) B ne peut s'attacher au-delà de la syllabe contenant la deuxième position nucléaire à partir de la fin du mot.

Comme on l'a souvent observé, de telles contraintes de placement ne font nulle allusion au poids (à la "quantité") syllabique, mais uniquement, dans (4), à la quantité des noyaux (une voyelle longue ou une diphtongue descendante étant représentée par deux positions nucléaires). On pourrait alors supposer que l'équivalent prosodique de la catégorie métrique de poids n'est pas, comme on l'a cru depuis toujours, le poids syllabique, mais bien la quantité des noyaux. Cette adaptation créerait hélas plus de problèmes qu'elle n'en résoudrait: elle nous obligerait soit à adopter une description fort curieuse pour le latin (où le poids est une catégorie prosodique), soit à renoncer au caractère présumé universel des équivalences dont la conjecture (1) postule l'existence.

Il n'est pas étonnant, dès lors, que divers auteurs aient tenté de montrer que la catégorie prosodique du poids jouait un rôle dans le système phonologique du grec ancien. Selon Saussure (1884) et bien d'autres (par exemple, Allen 1973: 52), le contraste entre σοφώτερος (< σοφός) d'une part, et ὠμότερος (< ὠμός) ou λεπτότερος (< λεπτός) d'autre part, s'explique, pour le premier de ces termes, par un allongement de /o/ à /ω/ qui est déclenché par la légèreté de la syllabe initiale. De même, la loi de Wheeler (5) et la loi de Vendryes (6) supposent la pertinence phonologique du poids (lourd ou léger) attribué à l'une ou l'autre syllabe:

(5) LOI DE WHEELER: lorsqu'un mot "oxyton" (portant H mais non B sur sa dernière syllabe) à voyelle finale brève contient une syllabe légère en position pénultième et une syllabe lourde en position antépénultième, alors H se déplace sur la pénultième. Exemple: *πᾰτρᾰσί > πατράσι.

(6) LOI DE VENDRYES: lorsqu'un mot porte la contonation HB sur sa syllabe pénultième (qui est dite à "accent circonflexe") et que sa syllabe antépénultième est légère, alors H se déplace sur l'antépénultième (qui est alors à "accent aigu"). Exemple: ἑτοῖμος > ἕτοιμος.

(voir Allen 1973: 204, 239; Lejeune 1955; Vendryes 1945).

De tels arguments se révèlent, en réalité, très fragiles. Tout d'abord, il a été soutenu (par exemple, par Kuryłowicz 1975) que les trois "processus" en question (la "loi rythmique" de Saussure, comme les lois de Wheeler et de Vendryes) doivent être remplacés par des explications d'un tout autre ordre. En outre, il s'agit à chaque fois de changements diachroniques dont on ne voit pas pourquoi ils possèderaient des corrélats à l'intérieur du système classique. Ainsi, la loi de Wheeler suppose, pour πᾰτρᾰσι, une syllabification πᾰτ-ρᾰσι ou πᾰτ-τρᾰσι attestée dans Homère, mais éliminée de la langue et de la métrique classiques par ce qu'on appelle traditionnellement la "correptio attica" (Allen 1973: 210-222).

Dans un article de 1988, Steriade a avancé un argument beaucoup plus intéressant en faveur de la pertinence synchronique du poids. Partant de l'observation bien connue selon laquelle les mots à finale consonantique complexe (se terminant sur un groupe obstruante + /s/) ne reçoivent jamais H ("l'accent aigu") sur l'antépénultième, Steriade soutient que le poids de la syllabe finale contraint le placement de l'accent. Je ne puis résumer ici le traitement, assez technique, que Steriade réserve à cette question. En gros, sa théorie prédit, par le jeu de l'extramétricité, que la syllabe finale sera légère à moins qu'elle ne contienne deux positions nucléaires ou qu'elle se termine sur un groupe obstruante + /s/. On peut donc remplacer le principe (3) par (3'):

(3') H ne peut s'attacher au-delà de la syllabe pénultième si la syllabe finale est lourde, et au-delà de la syllabe antépénultième si la syllabe finale est légère.

Par conséquent, si Steriade a raison, la conjecture (1) se voit confirmée pour ce qui concerne le statut métrique et prosodique du poids en grec ancien.

L'argument de Steriade apparaît d'autant plus décisif que nous ne disposions, pour ce qui concerne l'accentuation des mots à finales consonantiques complexes, que d'une explication peu convaincante initialement formulée par Bally (1945: 25-26, 111; cf. aussi Sommerstein 1973: 176-178), et sur laquelle je ne crois pas utile de m'étendre ici (cf. les critiques de Lupaş 1972: 33 et de Steriade 1988). Mais il est permis de se demander si le caractère marqué de mots comme φοῖνιξ ne déclenchait pas des syllabifications du type /pʰoi-nik-ks/ où la syllabe à laquelle s'ancre H se retrouve en position antépénultième. Une telle hypothèse simplifierait considérablement toute la description de l'accent grec, surtout si l'on doit prendre en compte les phénomènes d'enclise, que je n'ai pas le temps de commenter aujourd'hui.

Supposons, pour les besoins de notre réflexion, que cette hypothèse soit la bonne. Nous sommes alors revenus à notre point de départ. Nous ne savons toujours pas à quelle catégorie prosodique nous pourrions rattacher la catégorie métrique de poids et — chose plus grave encore — nous ne voyons pas clairement comment cette catégorie métrique a pu émerger. Les quelques processus diachroniques déjà évoqués sont trop limités, ou d'une existence trop douteuse, pour que nous puissions les exploiter ici sans autre inquiétude.

A ce stade, il nous faut recenser tous les phénomènes relativement centraux qui regroupent en une même classe les syllabes à noyaux complexes et certains types, au moins, de syllabes fermées. Pour autant que je sache, il n'existe qu'une donnée qui satisfasse à ces conditions; mais elle se révèle d'autant plus intéressante qu'elle concerne la langue homérique. Il existe en effet de bonnes raisons de croire que des syllabes contenant une sonante nasale ou liquide en position de coda pouvaient recevoir, comme les syllabes à noyau complexe, un accent dit "circonflexe", c'est-à-dire la totalité de la mélodie HB (Allen 1973: 242-243). En d'autres termes, la prosodie permettait de distinguer deux classes de syllabes: les syllabes "contonables" (susceptibles de porter un circonflexe), qui étaient nécessairement fermées par une sonante dans le cas ou leur noyau ne contenait qu'une position; et les syllabes "non contonables", à noyau syllabique simple, qui pouvaient être indifféremment ouvertes ou fermées. Il n'est pas interdit de penser qu'une métrique fondée sur cette catégorie prosodique a subi, suite à l'alignement des sonantes sur les autres consonnes, une réinterprétation qui a conduit à l'émergence métrique du poids.

L'adoption de cette reconstruction implique, bien évidemment, le rejet pur et simple de l'hypothèse (1). Il n'y aurait plus d'équivalence synchronique nécessaire entre catégories métriques et catégories prosodiques, mais simplement une exigence plus faible stipulant que l'association éventuelle de telle ou telle valeur métrique à telle ou telle unité linguistique soit toujours décidable, dans un sens positif ou négatif.

Prise telle quelle, cette formulation me paraît totalement insatisfaisante. Par exemple, elle n'interdit pas l'apparition d'une métrique quantitative qui opposerait les syllabes fermées aux syllabes ouvertes, indépendamment de toute pertinence de la quantité des noyaux. Pour bloquer pareille dérive, plusieurs solutions sont concevables a priori. Je serais tenté, pour ma part, de renoncer aux vues simplistes de certains métriciens, surtout français, qui se refusent à concevoir que l'accent ait pu jouer un rôle quelconque dans la métrique quantitative des Anciens (voir, par

exemple, Nougaret 1948). En termes plus précis, je serais enclin à croire qu'une métrique quantitative ne peut répartir les syllabes d'une langue en deux sous-classes complémentaires que s'il existe un ensemble de syllabes (par exemple les syllabes ouvertes) relativement auquel cette repartition se révèle déterminante pour la possibilité ou l'impossibilité de tel ou tel phénomène prosodique (ton modulé, accent circonflexe, etc.). Cette contrainte interdit, en grec ancien, l'utilisation d'une métrique quantitative uniquement sensible au caractère ouvert ou fermé des syllabes. En ce qui concerne le latin, où la dichotomie entre syllabes ouvertes et syllabes fermées se révèle déterminante, pour les syllabes à noyau simple, relativement au maintien ou au non-maintien de l'accent en position pénultième, un deuxième principe se trouve sans doute à l'œuvre. Je le formulerais comme suit: si le phénomène prosodique relativement auquel la distinction métrique s'avère déterminante permet de répartir toutes les syllabes de la langue en deux sous-classes complémentaires, alors la catégorie métrique de poids doit aboutir à la même partition.

Cet ultime développement de la discussion que je voulais consacrer à la métrique grecque m'amène à une autre question cruciale pour toute théorie de la versification. Nous avons pris l'habitude de caractériser les différentes métriques selon des typologies plus ou moins rigides (voir, par exemple, Lotz 1960 ou Jakobson 1960). Si les développements qui précèdent recèlent quelque vérité, ils devraient nous inciter à mesurer la véritable portée de ces idéalisations qui sont, par ailleurs, tout à fait légitimes. Un rapide examen de l'alexandrin français nous fournit des arguments allant dans le même sens.

Comme je l'ai rappelé en d'autres circonstances (Dominicy 1992), le "modèle de vers" de l'alexandrin racinien, comme de l'alexandrin "classique" en général, se divise en un premier sous-vers contenant 6 syllabes métriques et un second sous-vers contenant six syllabes métriques suivies d'une syllabe extramétrique:

(7) XXXXXX-XXXXXX(X)

Pour qu'une suite d'unités linguistiques S se revèle conforme, en tant qu'"exemple de vers", à ce patron sous-jacent, elle doit satisfaire à toute une série de conditions apparemment très dissemblables:

(8) S ne peut pas contenir d'hiatus (comme **tu es*);

(9) S ne peut pas contenir une séquence où un mot qui se termine par une voyelle suivie d'un "*e* muet" précède un mot dont la notation orthographique commence par une consonne (autre qu'"*h* muette");

(10) Lorqu'on a éliminé de S tous les "*e* muets" apparaissant à la fois en fin de mot et devant un mot commençant par une voyelle orthographique, on doit obtenir le nombre syllabique 13 si la dernière syllabe est posttonique, et le nombre syllabique 12 dans tous les autres cas;

(11) La syllabe 6 de S ne peut pas être posttonique;

(12) La syllabe 7 de S ne peut pas être posttonique;

(13) La syllabe 12 de S ne peut pas être posttonique;

(14) Les syllabes 6 et 7 de S ne peuvent pas appartenir au même mot;

(15) Si la syllabe 6 de S coïncide avec le début d'un constituant syntaxique qui s'étend (au moins) jusqu'à la syllabe 7, alors ce constituant est enchâssé dans un autre constituant (interne au vers) qui s'étend au moins jusqu'à la syllabe 5;

(16) La syllabe 6 ne peut coïncider avec un déterminant ou une préposition monosyllabique.

Les conditions (11) et (13) ont donné lieu à des débats, extrêmement confus, et qui se résumaient, pour l'essentiel, à la question de savoir si l'alexandrin français, et l'alexandrin racinien en particulier, est "accentuel" ou "non accentuel". En effet, compte tenu de la structure prosodique du français, la prohibition d'une syllabe posttonique en syllabes 6 et 12 entraîne que ces syllabes coïncideront soit avec un mot monosyllabique, soit avec la syllabe portant l'accent de mot primaire et principal à l'intérieur d'un mot plurisyllabique.

Du point de vue typologique, nous n'éprouvons guère de peine à trancher. Un modèle de vers est dit "accentuel" si, et seulement si, chacune de ses positions syllabiques reçoit l'une des deux qualifications "accentué" ou "non accentué". Comme toutes les syllabes autre que 6 et 12 se révèlent indifférentes, nous en concluons immédiatement que l'alexandrin n'est pas "accentuel". En d'autres termes, nous sommes bien forcés d'admettre qu'il peut exister de fortes corrélations entre certaines positions métriques et une certaine catégorie prosodique, sans que cela provoque nécessairement l'émergence de la catégorie métrique équivalente.

On a souvent soutenu, également, que la position 6 devait coïncider avec un "accent de groupe". Mais on trouve sans difficulté des vers de Racine qui infirment cette hypothèse:

(17) *J'y suis encor malgré tes infidélités*

(18) *Seigneur, si j'ai trouvé grâce devant vos yeux*

La seule généralité à laquelle on puisse arriver en cette matière découle des contraintes (11) et (15). En effet, elles aboutissent, prises conjointement, à exclure l'occurrence d'un accent de groupe sur la syllabe 5, et à favoriser l'apparition d'accents de groupe autour des syllabes 4 et 8. Une nouvelle fois, nous aboutissons à des corrélations prosodico-métriques qui ne conduisent pas à une situation d'équivalence entre catégories.

L'évolution de l'alexandrin, depuis Racine jusqu'à la fin du XIXe siècle, se caractérise par

l'apparition progressive et cumulative de certaines violations. J'en donne ici quelques exemples, rangés d'après la contrainte qu'ils transgressent:

(15) *Cette muraille, bloc d'obscurité funèbre*
(Hugo)

(16) *Ils s'en venaient de la montagne et de la plaine*
(Leconte de Lisle)

(11) *Nubiles plis l'astre mûri des lendemains*
(Mallarmé)

(12) *Dans votre sein, sur votre cœur qui fut le nôtre*
(Verlaine)

(14) *Cependant que, silencieux sous les pilastres*
(Rimbaud)

Grâce aux recherches de Cornulier (1982), nous savons que ces violations ont été compensées, très longtemps, par une prosodie du type 4-4-4, ou par l'un de ses dérivés (8-4 ou 4-8). Il est cependant permis de croire qu'à un certain stade, chez des auteurs comme Verlaine ou Mallarmé, ce processus a conduit à une réanalyse du modèle de vers initial. En d'autres termes, le modèle (7) se serait d'abord vu adjoindre le modèle ternaire (19), puis les modèles binaires (20) et (21):

(19) XXXX-XXXX-XXXX(X)
(20) XXXXXXXX-XXXX(X)
(21) XXXX-XXXXXXXX(X)

Si cette interprétation des données est correcte, elle montre qu'une catégorie prosodique qui est demeurée longtemps corrélée à des structures métriques peut, dans des circonstances qui resteraient à déterminer, induire une réorganisation des patrons sous-jacents. Ce type d'influence doit nous inciter à plus de prudence lorsque nous formulons des conjectures sur l'interaction entre le mètre et la prosodie, et sur les modalités que cette ineraction parfois très complexe peut revêtir.

RÉFÉRENCES

ALLEN, W.S. (1973), *Accent and Rhythm. Prosodic Features of Latin and Greek: a Study in Theory and Reconstruction*, Cambridge, University Press.

BALLY, C. (1945), *Manuel d'accentuation grecque*, Berne, Francke.

CORNULIER, B. de (1982), *Théorie du vers. Rimbaud, Verlaine, Mallarmé*, Paris, Editions du Seuil.

DOMINICY, M. (1992), "On the Meter and Prosody of French 12-Syllable Verse", à paraître dans Grimaud, M. (éd.), *Foundations of Verse* (special issue of *Empirical Studies of the Arts*).

JAKOBSON, R. (1960), "Closing Statement: Linguistics and Poetics", dans Sebeok, 350-377.

KURYŁOWICZ, J. (1975), "De σοφώτερος à δύσερως", dans *Mélanges offerts à Emile Benveniste*, Paris, Klincksieck, 325-330.

LEJEUNE, M. (1955), *Traité de phonétique grecque*, Paris, Klincksieck, deuxième édition.

LOTZ, J. (1960), "Metric Typology", dans Sebeok, 135-148.

LUPAŞ, L. (1972), *Phonologie du grec attique*, La Haye-Paris, Mouton.

NOUGARET, L. (1948), *Traité de métrique latine classique*, Paris, Klincksieck.

SAUSSURE, F. de (1884), "Une loi rythmique de la langue grecque", repris dans le *Recueil des publications scientifiques*, Heidelberg, Winter, 1922, 464-476.

SEBEOK, T.A. (1960), éd., *Style in Language*, Cambridge (Mass.), The M.I.T. Press.

SOMMERSTEIN, A.H. (1973), *The Sound Pattern of Ancient Greek*, Oxford, Blackwell.

STERIADE, D. (1988), "Greek accent: a case for preserving structure", *Linguistic Inquiry* 19, 271-314.

VENDRYES, J. (1945), *Traité d'accentuation grecque*, Paris, Klincksieck, nouveau tirage.

VERLUYTEN, S.P. (1989), "L'analyse de l'alexandrin. Mètre ou rythme?", dans Dominicy, M. (éd.), *Le souci des apparences*, Bruxelles, Editions de l'Université, 31-74.

ZIRIN, R.A. (1970), *The Phonological Basis of Latin Prosody*, La Haye-Paris, Mouton.

19es J.E.P. - Bruxelles, 19 au 22 mai 1992.

ANALYSE DE LA PRODUCTION DES VOYELLES DE QUELQUES LANGUES SOUDAN CENTRAL PAR IRM.

DIDIER DEMOLIN - CHRISTOPH SEGEBARTH

UNIVERSITE LIBRE DE BRUXELLES
SERVICE DE LINGUISTIQUE GENERALE - HOPITAL ERASME

Résumé

Les techniques d'imagerie par résonance magnétique ont été utilisées pour étudier les configurations articulatoires des voyelles de trois langues soudan central parlées au Zaïre. Nous pouvons ainsi montrer que le position de la racine de la langue (trait [±ATR]) est bien un paramètre déterminant pour distinguer les voyelles [i/ɪ, e/ɛ, o/ɔ, u/ʊ]. Les images obtenues avec cette technique permettent aussi de montrer que les voyelles centrales de ces langues se caractérisent par une position assez haute dans le conduit oral. En lendu, une des voyelles centrales ([ɐ]) est nasalisée, chez les sujets qui ont participé à l'étude.

1. INTRODUCTION.

Les techniques d'imagerie par résonance magnétique (dorénavant IRM) ont été utilisées pour obtenir les configurations articulatoires du conduit vocal avec des locuteurs produisant des voyelles soutenues. Ces techniques sont particulièrement utiles pour obtenir, d'une manière non-envahissante et inoffensive, les dimensions de la région pharyngale. Cette investigation a été réalisée à l'Unité de Résonance Magnétique de l'Hôpital Erasme, de l'Université Libre de Bruxelles. Nous avons employé un imageur Gyroscan S15 (Philips, 1.5 Tesla). Les mesures ont été faites avec l'antenne RF "cerveau", utilisée pour l'excitation RF ainsi que pour la détection des signaux nucléaires. La technique RF est une technique spin-écho (SE) avec un temps de répétition de 200 ms (le minimum réalisable sur la machine); le temps d'écho (TE) est de 25 ms, le champ de vision de 250 mm, et la matrice d'acquisition de 128*256.

Dans un premier examen, on a procédé à un repérage de 10 coupes coronales adjacentes de 10 mm d'épaisseur, dont le temps d'acquisition total est de 1.53 minute. L'examen proprement dit a consisté en une coupe médio-sagittale de 8 mm d'épaisseur, dont le temps d'acquisition est de 19 secondes.

Dans cet article, nous présentons une étude articulatoire de la production des voyelles du mangbetu, du lendu et du lugbara.

Nous avons étudié la réalisation des voyelles de chaque langue avec un ou deux locuteurs. Notre but est d'établir la pertinence véritable du trait ATR, qui est employé à la fois pour caractériser les voyelles et pour décrire le processus d'harmonie vocalique, dans les langues de cette famille linguistique. Un deuxième objectif est d'observer la réalisation des voyelles ouvertes dans les trois langues ainsi que les voyelles centrales du lendu.

L'étude vise essentiellement à caractériser les paramètres articulatoires qui sont pertinents du point de vue linguistique, et non à quantifier les résultats pour produire un modèle mathématique[1].

La clarté des documents (cf. les 9 clichés de la figure 1, pour un sujet masculin mangbetu) permet de caractériser les paramètres articulatoires qui entrent en jeu dans la réalisation des voyelles, à savoir:

(i) la protusion des lèvres , qui ne doit pas être confondue avec l'ouverture des lèvres , puisque celles-ci sont vues en coupe médio-sagittale;
(ii) la position de la pointe de la langue , du dos de la langue , de la masse de la langue et de la racine de la langue ;
(iii) la position du velum;
(iv) l'abaissement de la mâchoire et la rétraction de la mâchoire ;
(v) la position du larynx.

Ceci dit, la technique IRM utilisée offre malgré tout certains inconvénients.

---

[1] Cette étude présente les résultats préliminaires d' un programme de recherche plus vaste qui cherche à établir un modèle de la production des voyelles dans les langues soudan central. Le nombre limité d'enregistrements et le peu de locuteurs étudiés ne nous permet pas de produire de résultats statistiquement significatifs.

Le fait que les voyelles doivent être soutenues pendant des périodes assez longues peut introduire des gestes de compensation, ou forcer le mouvement des articulateurs, sans que cela soit décelable. D'autre part, le bruit de la machine interdit tout contrôle auditif pendant l'émission des voyelles.

Les structures calcifiées (os et dents) sont parfois difficiles à identifier avec précision. Les dents, qui contiennent peu d'hydrogène mobile, sont totalement indiscernables du conduit lors de l'émission des voyelles.

Les contraintes que l'étude impose sur les temps d'enregistrements (une voyelle peut difficilement être soutenue pendant plus de 25 s) entraînent une diminution de la qualité de résolution des images. La résolution des limites air-tissus dépend aussi de l'épaisseur de la section de tissus qui génère le signal de résonance magnétique. Pour obtenir des images de qualité suffisante, nous avons fixé l'épaisseur des coupes à 8 mm.

On peut également se demander si la position couchée dans laquelle le sujet reste pour une période assez longue n'entraîne pas un effet de gravitation affectant les organes de la parole. Ce problème a déjà été mentionné par Baer, Gore, Gracco et Nye (1991), mais jusqu'à présent, aucun effet mesurable dû à la gravitation n'a encore été observé. Au cours des expériences (dont la durée allait jusqu'à 3 heures), l'ordre d'enregistrement des voyelles était [i, ɪ, e, ɛ, a, ɔ, o, ʊ, u], ou l'inverse. La comparaison des voyelles antérieures ou postérieures enregistrées au début et à la fin des séances ne présente aucune différence mesurable due à un effet de gravitation. Foldvik et al. (1991 et communication personnelle) ont fait des constatations similaires à propos des voyelles du norvégien.

Malgré ces quelques inconvénients, une description articulatoire fondée sur l'IRM permet de caractériser avec précision les configurations articulatoires de la production des voyelles. Outre les coupes médio-sagittales, il serait utile de recourir à des coupes transversales, afin de mieux caractériser le volume du pharynx, et afin de développer des modèles à trois dimensions du mouvement de la langue (Stone 1991). Cependant, comme l'étude qui suit vise à déterminer les configurations articulatoires, il n'a pas été tenu compte de cet aspect.

2. METHODE.

Chaque voyelle, réalisée à cinq reprises au cours d'une séance, était soutenue pendant une période de 19 ou 25 s. Les résultats des enregistrements IRM ont été transférés sur des films radiographiques qui ont permis de tracer les contours articulatoires de chacune des voyelles. Le but de cette investigation était d'observer et de comparer la réalisation de chaque voyelle du système chez les sujets, en se concentrant sur deux points en particulier: le mouvement de la racine de la langue et la position de la masse de la langue.

La grille d'analyse (figure 2), qui permet la mesure et la comparaison de chacune des voyelles, est constituée de la manière suivante. Sur chacun des tracés, les contours de l'os du palais dur, de l'os sphénoïde, de l'extrémité inférieure de l'os occipital, et des quatre premières vertèbres, sont dessinés. Ces structures osseuses apparaissent en noir sur les clichés, à l'exception de la moelle et de la graisse qui apparaissent en blanc. Leur superposition permet de comparer la réalisation de chaque voyelle, et de dessiner le contour des articulateurs. La grille elle-même est élaborée en traçant une droite SX qui va de l'extrémité inférieure de l'os sphénoïde à l'extrémité inférieure de l'os occipital. La distance entre les deux points de contact S et X est constante chez les sujets observés. Une droite parallèle à SX, S'X', est ensuite tracée de manière à passer au centre de la langue. La distance SX est donc reportée sur S'X'. A une distance de 4.5 cm de S', vers le centre de la langue, une droite perpendiculaire coupant S'X' au point O est tracée. Tous les 15° par rapport au plan S'X', et en passant par le point O, une droite est tracée. Le déplacement des articulateurs peut alors être reporté sur chaque droite qui rayonne à partir de O. Le contour moyen de chaque voyelle est tracé à partir des valeurs moyennes du déplacement des articulateurs, mesuré sur les droites qui passent par O.

3. VOYELLES MANGBETU.

Le système vocalique du mangbetu comporte 9 voyelles [i, ɪ, e, ɛ, a, ɔ, o, ʊ, u] (cf. Demolin 1992).

3.1. VOYELLES ANTERIEURES [i, ɪ, e, ɛ].

La comparaison globale des voyelles antérieures (figure 3) montre que le mouvement de la racine de la langue est bien un paramètre pertinent dans la production des voyelles antérieures du mangbetu. Ce mouvement est cependant beaucoup moins diversifié que celui qui a été décrit par Stewart (1967) et Lindau (1979) en akan, et par Ladefoged (1964) en igbo. Le mouvement de rétraction de la racine de la langue et du phrarynx lors du passage de [i] à [ɛ] est le plus marqué à la hauteur de l'épiglotte. La position du larynx ne paraît pas varier beaucoup, sauf en ce qui concerne [ɛ], où il est légèrement plus bas.

Deux autres faits sont à remarquer à propos des voyelles antérieures. Le premier est la présence d'une sorte de plateau articulatoire situé dans la région palatale. C'est une zone où le contour du dos de la langue est dans une position stable pour toutes les voyelles. L'écart entre la voyelle la plus fermée [i] et la voyelle la plus ouverte [ɛ] est assez minime et constant dans cette zone. Le second fait est le mouvement d'abaissement de la mâchoire inférieure qui, combiné à la position de la pointe de la langue, conditionne le volume de la cavité antérieure. L'ordre décroissant de cette ouverture maxillaire est: [e, i, ɪ, ɛ]. Le

volume de la cavité antérieure conditionne la réalisation acoustique de F2, dont la fréquence décroît quand on passe de [i] à [ɛ]. Il est obtenu en combinant l'abaissement de la mâchoire et un ajustement de la partie antérieure de la langue; un seul de ces deux paramètres ne suffit apparemment pas, chez les locuteurs observés

La comparaison deux à deux des voyelles antérieures montre clairement les différences entre [i/ɪ] et [e/ɛ]. Les différences les plus sensibles se situent au niveau du volume de la cavité pharyngale pour les voyelles les plus fermées, et au niveau du volume de la cavité antérieure pour les voyelles les plus ouvertes.

La comparaison des voyelles [ɪ] et [e] indique une différence bien marquée dans la position de la racine de la langue, ainsi qu'une position sensiblement identique de la cavité antérieure, malgré une fermeture légèrement plus grande pour [ɪ].

3.2. VOYELLES POSTERIEURES [ɔ, o, ʊ, u].

La comparaison globale des voyelles postérieures (figure 4) montre également que le mouvement de la racine de la langue est un paramètre essentiel dans la production de ces voyelles. La différence entre [ɔ, o, ʊ, u] est très nette, et bien plus marquée que pour les voyelles antérieures, au niveau de la racine de la langue. La rétraction de la racine de la langue se fait dans l'ordre suivant: [u, o, ʊ, ɔ]. La position du larynx varie légèrement d'une voyelle à l'autre. Le niveau d'abaissement suit l'ordre [u, o, ʊ, ɔ], mais les écarts paraissent peu significatifs.

La position et le mouvement de la mâchoire sont très différents de ce qu'on observe avec les voyelles antérieures. En plus du mouvement d'abaissement de la mâchoire, il y a un mouvement de rétraction nettement perceptible. La combinaison de ces deux paramètres semble assez complexe. L'abaissement observé à l'ouverture de la bouche se fait dans l'ordre [u, o, ʊ, ɔ], mais si l'on regarde le profil inférieur de la mâchoire, l'ordre d'abaissement devient [o, u, ʊ, ɔ]. Cette combinaison de l'abaissement et de la rétraction de la mâchoire inférieure paraît également jouer un rôle déterminant sur la forme de la racine de la langue, et sur le volume du pharynx.

Le rôle du larynx, et en particulier du mouvement indépendant qu'il pourrait avoir par rapport au mouvement de la racine de la langue, est difficile à évaluer sur la base de l'information fournie par le type de données que nous observons.

La position de la partie antérieure de la langue présente aussi quelques aspects particuliers. Stewart (1967) et Lindau (1979) ont décelé, pour cette partie antérieure, un abaissement progressif qui va de [u] à [ɔ] en passant par [ʊ] et [o]. Chez le sujet observé, la forme de la langue paraît varier beaucoup si on compare les quatre voyelles postérieures. En comparant ces voyelles deux à deux on constate que [o] et [ɔ] ont des formes similaires, et que la masse de la langue est plus élevée pour [o], ce qui était prévisible. En ce qui concerne les voyelles fermées [u, ʊ], la position de la langue paraît moins évidente à interpréter. La hauteur de la masse de la langue est identique pour ces deux voyelles dans la région postérieure, mais la partie antérieure de la langue de [ʊ] est beaucoup plus avancée, et a une forme assez différente de celle de [u]. La forme de [ʊ] fait penser à une voyelle haute plutôt centralisée. Ce phénomène confirme le caractère centralisé attribué à [ʊ] (cf. Demolin 1992). La comparaison de [ʊ] et [o] montre que ces deux voyelles, qui se distinguent par la position de la racine de la langue, présentent toutes deux une sorte de plateau articulatoire situé au niveau de la masse de la langue. La pointe de la langue est cependant dans une position plus reculée pour [ʊ].

3.3. VOYELLE [a].

La voyelle [a] (figure 5) présente un contour assez inattendu, vu la forme de la langue qui est beaucoup plus haute que ce que l'on aurait pu prévoir. Pour vérifier et contrôler cet aspect de la réalisation de [a], nous avons quadruplé les enregistrements de cette voyelle au cours des séances d'enregistrements, sans observer de variation significative dans sa réalisation. On peut observer quelques faits intéressants en comparant [a] à [ɔ] et à [ɛ] (figure 6). L'aperture et la forme des lèvres sont assez comparables entre [a] et [ɛ]. La voyelle postérieure [ɔ] a une protrusion des lèvres nettement marquée quand on la compare aux deux autres voyelles. Les trois voyelles ont une cavité pharyngale dont le volume est assez réduit. Le larynx est le plus bas avec [ɔ]. La voyelle [ɛ] est la plus avancée, tandis que [a] et [ɔ] sont similaires sur ce point. Ce qui distingue nettement ces voyelles, c'est la forme de la masse de la langue. La pointe — et non la masse — de la langue est la plus rétractée et la plus basse pour [a].

4. VOYELLES LENDU.

Le système vocalique du lendu comporte 8 voyelles [i, ɪ, ɛ, ɵ, a, ɔ, ʊ, u] (cf. Lojenga 1989).

4.1. VOYELLES ANTERIEURES.

La comparaison des trois voyelles antérieures (figure 7) montre un léger mouvement de la racine de langue entre [i, ɪ, ɛ]. La différence entre les deux voyelles fermées [i/ɪ] paraît la mieux marquée entre l'épiglotte et le dos de la langue. La différence d'aperture

entre les deux voyelles fermées n'est pas très sensible. Il semble donc que la position de la racine de la langue soit déterminante pour leur distinction.

4.2. VOYELLES POSTERIEURES.

Les voyelles postérieures du lendu (figure 8) indiquent aussi qu'il y a un mouvement significatif de la racine de la langue qui intervient dans la production de ces voyelles. Comme en mangbetu, la position et le mouvement de la mâchoire sont très différents de ce qu'on observe avec les voyelles antérieures. Ici aussi, l'abaissement et la rétraction importante de la mâchoire semblent jouer un rôle déterminant sur la forme de la racine de la langue te sur le voulme du pharynx. La position de la partie antérieure de la langue s'abaisse progressivement de [u] à [ɔ] en passant par [ʊ].

La position du larynx ne paraît pas varier de manière significative.

4.3. VOYELLES [ɐ et ɵ].

Le lendu a dans son inventaire vocalique une voyelle [ɵ] dont le statut phonétique et phonologique à fait l'objet de discussions importantes parmi les descripteurs de cette langue (cf. Lojenga 1989 pour un résumé). Cette voyelle a été successivement décrite comme: antérieure arrondie [ø], postérieure fermée et centralisée [ɯ], et centrale non-fermée [+ATR] [ɵ]. L'observation des contours obtenus à partir des coupes médio-sagittales en IRM montre que chez les deux locuteurs dont nous avons observé la production, cette voyelle peut se définir comme centrale non-fermée [+ATR] et nasalisée. L'avancement de la racine de la langue est particulièrement visible si on compare cette voyelle avec [ɐ] qui est ouverte centrale et [-ATR] (figure 9). La voyelle [ɵ] se caractérise aussi par un avancement plus important de la partie antérieure de la langue et par une aperture plus petite que celle de [ɐ].

5. LUGBARA.

L'examen des données lugbara indique que le trait [±ATR] semble aussi pertinent pour distinguer les paires de voyelles [i/ɪ], [u/ʊ] et [o/ɔ]. Ces données sont toutefois problématiques à cause de la résolution assez pauvre des images que nous avons obtenues et par la configuration aberrante de certaines voyelles (contact avec la zone palatale pendant l'enregistrement). C'est la raison pour laquelle elles ne sont pas présentées ici.

6. CONCLUSION.

L'articulation des voyelles mangbetu se caractérise, chez le sujet observé, par des propriétés assez différentes de celles qui ont été décrites dans les études cinéradiographiques portant sur d'autres langues africaines pourvues de systèmes vocaliques similaires. Il est cependant difficile de tirer des conclusions générales quant au système du mangbetu à partir de l'observation d'un seul locuteur. En effet, comme on l'a souvent souligné (cf., par exemple, Lindau 1979: 164), les locuteurs d'une même langue peuvent parfaitement réaliser des voyelles acoustiquement similaires au moyen de gestes articulatoires différents.

Les résultats obtenus en lendu montrent que chez les deux sujets observés, la position de la racine de la langue est un paramètre important dans la production des voyelles. La configuration de la voyelle [ɵ] confirme la description de Lojenga (1989), mais elle introduit un nouveau paramètre qui est la nasalisation de cette voyelle. Il convient toutefois de vérifier cette donnée avec un nombre plus important de locuteurs pour la confirmer et savoir si elle n'est pas due à une déviation provoquée par l'expérience.

REFERENCES

Baer, T. J.C. Gore, S. Boyce et P.W. Nye. (1987) "Application of MRI to the analysis of speech production", *Magnetic Resonance Imaging* 5, 1-7.

Baer, T., J.C. Gore, L.C. Gracco et P. W. Nye. (1991) "Analysis of vocal tract shape and dimensions using magnetic resonance imaging: vowels", *Journal of the Acoustical Society of America* 90, 879-828.

Demolin, D. (1992) *Le mangbetu: étude phonétique et phonologique*, thèse de doctorat, Université libre de Bruxelles.

Foldvik, A.K., O. Husby, J. Kvaerness, I.C. Norli et P.A. Rinck. (1991) "MRI (magnetic resonance imaging) for filming articulatory movements", *Proceedings of the 12th ICPhS*, Aix-en-Provence, 34-36.

Ladefoged, P. (1964) *A Phonetic Study of West African Languages*, Cambridge university Press.

Lindau, M. (1979) "The feature expanded", *Journal of Phonetics* 7, 163-176.

Lojenga, C. (1989) "The secret behind vowelless syllables in Lendu", *Journal of African Languages and Linguistics* 11, 115-126.

Stewart, J. (1967) "Tongue root position in Akan vowel harmony", *Phonetica* 16, 185-204.

Stone, M. (1991) "Toward a model of three-dimensional tongue movement", *Journal of Phonetics* 19, 309-320.

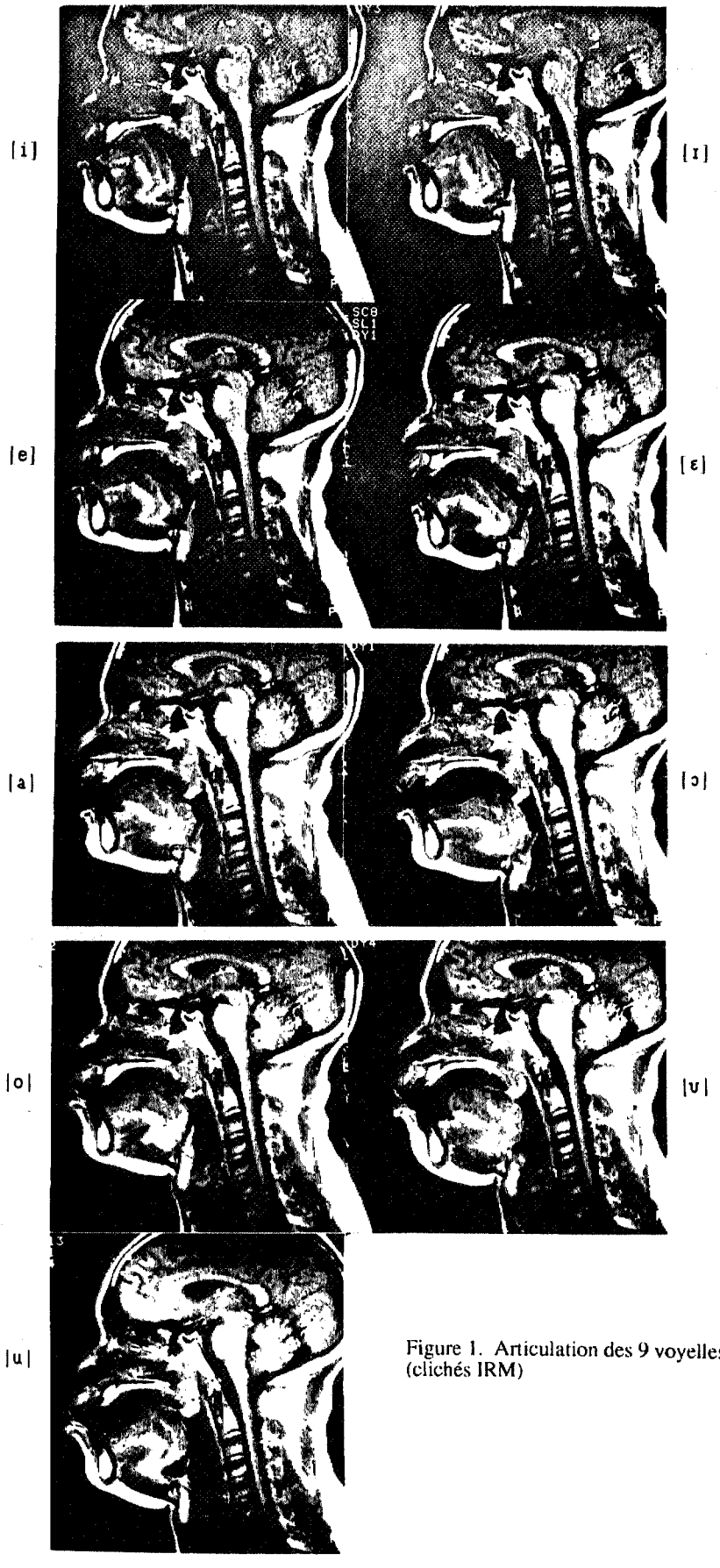
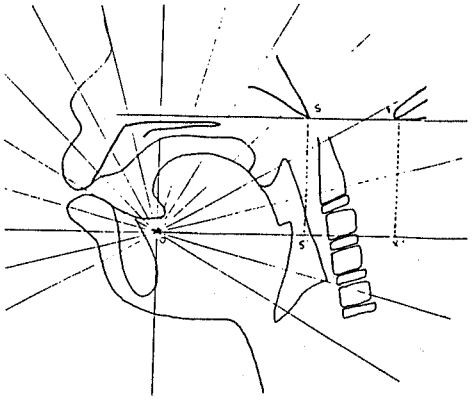
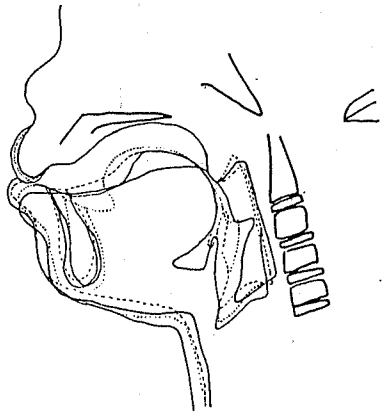


Figure 1. Articulation des 9 voyelles mangbetu (clichés IRM)

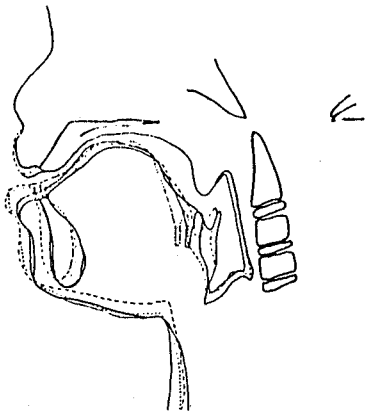
Figure 2. Grille d'analyse des clichés IRM





[u] ——
[ʊ] ·····
[o] - - -
[ɔ] ——

Figure 4. Voyelles postérieures mangbetu



[i] ·····
[ɪ] ——
[e] ·····
[ɛ] ——

Figure 3. Voyelles antérieures mangbetu

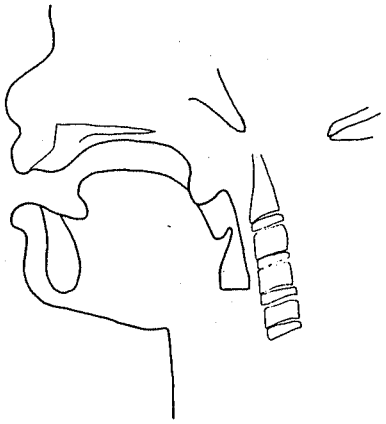
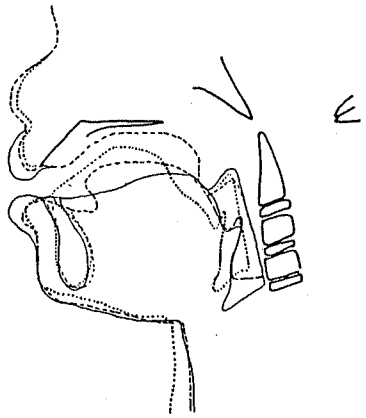
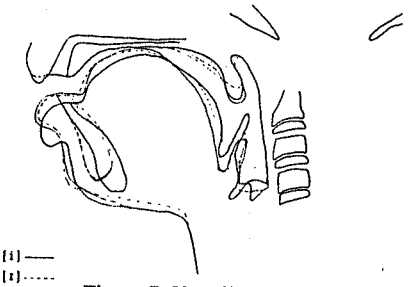


Figure 5. Voyelle [a] mangbetu



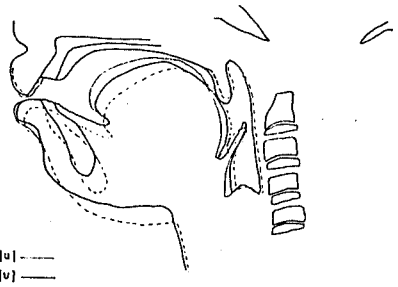
[a] ------
[ɛ] ·····
[ɔ] ——

Figure 6. Voyelles [a / ɛ / ɔ] mangbetu



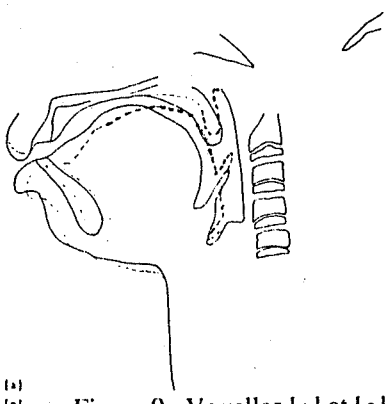
[i] ——
[ɪ] - - -
[ɛ] ——

Figure 7. Voyelles antérieures lendu



[u] - - -
[ʊ] ——
[ɔ] ------

Figure 8. Voyelles postérieures lendu



[a]
[ɵ] —

Figure 9. Voyelles [a] et [ɵ] lendu

DE LA DIFFICULTE DE COMPARER LES SYSTEMES VOCALIQUES

Jean-Marie HOMBERT

LAPHOLIA
Université Lumière-Lyon 2

L'étude des universaux des systèmes vocaliques implique des choix concernant les langues, les sources bibliographiques, les éléments représentés, et les types de représentation . Nous examinons les problèmes soulevés par ces différents choix ainsi que par les modèles utilisés pour tester les principes qui sous-tendent ces universaux. Enfin, nous proposons trois axes de recherche susceptibles de faire progresser notre compréhension du nombre et de la distribution des voyelles d'un système donné.

1 - CONSTITUTION DU CORPUS

Toute étude sur les universaux repose sur la constitution d'un corpus représentatif. La détermination de ce corpus implique des choix de langues, de références bibliographiques et de types d'analyses linguistiques.

1-1 Choix des langues

Les langues retenues doivent être représentatives de l'ensemble des groupes et sous-groupes linguistiques. Malheureusement, pour la plupart des familles de langues du monde -en particulier pour les langues à tradition orale- cette classification linguistique est tout à fait préliminaire. Il est par conséquent très difficile de choisir des langues représentatives en particulier au niveau des sous-groupes.

1-2 Choix des sources

Lorsque plusieurs références bibliographiques existent pour la même langue, il faut alors choisir entre une tentative de mise en commun des informations linguistiques provenant de ces différentes sources (avec tous les risques de confusion des variations dialectales) ou de ne retenir qu'une seule source considérée comme représentative. En outre, les analyses phonologiques peuvent avoir des niveaux très différents (très abstraites ou très proches de la réalité phonétique) ; il s'agit alors, si l'on souhaite pouvoir comparer les

analyses des différentes langues, d'uniformiser les "niveaux d'analyse".

1-3 Choix des éléments représentés

1-3-1 Phonèmes vs. allophones

Certains "sons" ne se trouvent que dans des contextes bien précis, parfois prévisibles (on peut alors parler d'allophones) ; faut-il dans ce cas les exclure de l'inventaire phonique de la langue considérée même s'ils occupent un "trou" dans cet inventaire ?

1-3-2 Séries secondaires

On considère généralement comme série primaire dans les inventaires des systèmes vocaliques la série des voyelles orales brèves. Toutefois dans certaines langues, la série de voyelles longues représente la série de base : par exemple en Tigre (langue sémitique) la série de voyelles longues est constituée par i:, e:, a:, o:, u: alors que la série de voyelles brèves n'est constituée que par e et œ.

Les séries secondaires quant à elles sont principalement constituées et/ou par les séries nasales et les séries pharyngalisées. Lorsque les timbres des voyelles de ces séries ne sont pas identiques aux timbres des voyelles de la série primaire, faut-il alors les prendre en compte ? Faut-il ne prendre en compte que celles qui sont distinctes ? Pour illustrer notre propos, prenons le cas d'une langue hypothétique ayant comme voyelles i, æ, a, o, u / i:, e:, a:, ɔ:, u: / i̪, I̪, a̪, U̪, u̪ · faut-il considérer cette langue comme ayant un système à 5, 7, 9 ou 15 voyelles ?

Un problème supplémentaire est posé par les diphtongues : faut-il ou non les inclure dans l'inventaire des voyelles ? Dans l'un des parlers Fang (langue Bantu du Nord Gabon) on trouve la diphtongue wa mais pas la monophtongue ɔ (alors que dans les parlers voisins, on trouve ɔ et non pas wa) ; un inventaire des voyelles orales de ce parler ne prenant en considération que les monophtongues ferait apparaître un "trou" dans la région du ɔ.

Enfin, le phénomène d'harmonie vocalique pose un problème complexe: en effet, lorsque seulement certains timbres des voyelles constituant l'ensemble du système vocalique peuvent se réaliser dans une position donnée - à cause des contraintes imposées par le processus d'harmonie vocalique- est-il justifier de prendre en compte l'ensemble des timbres du système vocalique comme représentatif de la langue. Plus précisément, les langues possédant une harmonie de type ATR telles que le Diola, le Dan, l'Amo, le Zandé, le Luo qui possèdent un inventaire complet contenant entre 8 et 10 voyelles (voir par exemple l'inventaire UPSID (7) devraient être analysées comme système à 5 voyelles (avec dédoublement total ou partiel) puisque la contrainte d'harmonie vocalique implique qu'à l'intérieur du domaine d'harmonie - en général ,le mot - seulement 5 timbres différents sont possibles.

Le même type de commentaire s'applique aux langues du groupe Ouralo-altaïque qui sont sujettes à une forme d'harmonie vocalique basée soit sur l'arrondissement, soit sur le lieu d'articulation (8).

2 - REPRESENTATION DES SYSTEMES VOCALIQUES

Le premier obstacle rencontré lors de la constitution d'une base de données sur les systèmes vocaliques est le manque de données précises. On ne dispose en effet de données expérimentales (articulatoires et

acoustiques) que pour un nombre très restreint de langues.

Même pour les langues pour lesquelles ces données sont disponibles, trois types de décisions influent sur le mode de représentation des systèmes vocaliques :

2-1 Choix des paramètres représentés

Les formants sont les paramètres le plus fréquemment choisis ; toutefois la prise en compte de l'ensemble du spectre peut être jugée préférable (6).

2-2 Nombre de dimensions de l'espace ;

Lorsque l'on ne dispose pas de données expérimentales on est contraint d'utiliser un espace à deux dimensions représentant l'aperture et le lieu d'articulation. Avec les données acoustiques, on peut se limiter à ces deux dimensions (qui représenteront alors le premier et le second formant) ou même prendre en compte les formants supérieurs soit à l'intérieur de la deuxième dimension (mesures de type F'2), soit en augmentant le nombre de dimensions.

2-3 Domaine de la représentation

Dans quelques rares cas, ou les données perceptuelles existent on pourra utiliser une représentation dans le domaine perceptuel et non pas dans le domaine acoustique plus ou moins modifié par des considérations psychoacoustiques.

3 - MODELISATION

Après avoir sélectionné un corpus représentatif et un type de représentation, on peut alors tester les hypothèses permettant de prédire les différents types de systèmes vocaliques observés. Pour cela, il est nécessaire de définir un modèle permettant de tester ces hypothèses. Outre le choix du niveau de représentation évoqué au paragraphe précédent, deux types de décisions seront centrales au modèle : le type de distance utilisé et la nature de l'optimisation. Sur ce dernier point, les premiers modèles étaient fondés sur la recherche d'une distance maximale entre les différents éléments du système vocalique (5), alors que les modèles plus récents optent plutôt pour une "distance suffisante".

4 - NOUVELLES ORIENTATIONS

Il nous semble évident que l'étude des universaux des systèmes vocaliques et des modélisations permettant de tester les principes qui régissent ces universaux, en particulier concernant le nombre et la distribution des voyelles à l'intérieur du système, nous permet de mieux éclairer les processus de codage et décodage de la parole. Les travaux les plus récents dans ce domaine (6, 7, 9) ont exploité au maximum les corpus synchroniques existants. Pour progresser dans notre compréhension de ces universaux, trois axes au moins nous semble prioritaires.

4-1 Constitution d'un corpus "normalisé"

Nous avons rappelé au paragraphe 2 que l'une des difficultés rencontrée dans la constitution de corpus de systèmes vocaliques était le manque de données précises. Il faut aussi ajouter que même lorsque de telles données existent elles ne sont pas toujours facilement comparables : le problème de la normalisation des mesures formantiques n'est pas résolu. Une solution possible serait de constituer une base de données

perceptuelles à partir d'un même ensemble de stimuli vocaliques synthétiques (2,4). En procédant ainsi on évite des décisions arbitraires sous-jacentes à tous systèmes de normalisation automatique puisque cette normalisation est effectuée par les sujets eux-mêmes.

4-2 Prise en compte des exceptions

L'examen des corpus de systèmes vocaliques permet de dégager des tendances universelles quant au nombre et à la distribution des voyelles à l'intérieur du système. La prise en compte des systèmes qui sont hors de ces tendances universelles peut, elle aussi, être riche d'enseignement. Nous pensons par exemple aussi bien au rôle des processus d'harmonie vocalique mentionné au paragraphe 1-3-2 (à noter l'importance du nombre de voyelles antérieures arrondies dans les langues Ouralo-altaïques) (7,8) qu'au statut marginal de certains phonèmes résultant soit d'emprunts, soit de réalisations phonétiques de surface (voir par exemple le schwa dans les parlers berbères ou arabes).

4-3 Prise en compte de la diachronie

La prise en compte de la diachronie permettrait d'éclairer deux aspects non étudiés dans les analyses des systèmes synchroniques :

- Les conditions de passage d'un système de n voyelles à un système à m voyelles (m > ou < que n).

- Les critères de choix entre deux systèmes possibles attestés en synchronie (par exemple les deux types de systèmes à quatre ou à six voyelles) en fonction du type de segments (ou de classes de segments) qui a déclenché ce changement. Nous faisons allusion ici, par exemple, à l'antériorisation de voyelles postérieures déclenchée par les consonnes dentales (3).

REFERENCES :

(1) - HOMBERT J.-M., 1977 : "A model of tone systems", *Working Papers in Phonetics, UCLA* 36, 20-32.

(2) - HOMBERT J.-M., 1979 : "Universals of vowel systems : the case of centralized vowels", *Proceedings of 9th International Congress of Phonetic Sciences*, vol. 2, Copenhagen,

(3) - HOMBERT J.-M., 1984 : *Phonétique Expérimentale et Diachronie : Application à la Tonogénèse* - Thèse d'Etat, Université de Provence, 2 v.

(4) - HOMBERT J.-M. et G. PUECH, 1984 : "Espace Vocalique et Structuration Perceptuelle : Application au Swahili", *Pholia 1*, 199-208

(5)- LILJENCRANTS J. et LINDBLOM B., 1972 : "Numerical Simulation of Vowel Quality Systems : The Role of Perceptual Contrast", *Language 48*, 839-862.

(6) - LINDBLOM B., 1986 : "Phonetic Universals in Vowel Systems" in *Experimental Phonology* edited by J. J. OHALA and J. J. JAEGER, Academic Press, 13-44.

(7) - MADDIESON I., 1984 : *Patterns of Sounds*, Cambridge University Press.

(8) - VAGO R., 1980 : *Issues in Vowel Harmony*, John BENJAMINS, Amsterdam.

(9) - VALLEE N., L. J. BOE et J. L. SCHWARTZ, 1991 "Tendances Universelles et Stabilité des Systèmes Vocaliques", *Actes du XIIème Congres International des Sciences Phonétiques - Aix-en Provence*, 142-145.

19es J.E.P. - Bruxelles, 19 au 22 mai 1992.

UNE AUTRE VISION DES MODÈLES D'ANTICIPATION

Mohamed Tahar LALLOUACHE et Christian ABRY

**Institut de la Communication Parlée, U.A. CNRS N° 368,
INPG/ENSERG - Université Stendhal, BP 25X - 38040 Grenoble, France**

Résumé

Entre deux modèles d'anticipation – dits *look-ahead* [LA] et *time-locked* [TL] – qui s'affrontent sur le *trait/geste* d'arrondissement depuis des années, le débat s'est focalisé au bout du compte sur l'explication de signaux «hybrides», comprenant une première phase lente expliquée, soit par la présence de gestes consonantiques de protrusion [s, t, l, ...] (*sic*) en compétition avec le geste vocalique (version *coproduction*, BOYCE *et al.*, 1991); soit (et/ou) par la compétition entre : la tendance à anticiper aussi tôt que possible [LA] et la tendance à garder la durée du mouvement constante [TL] (PERKELL et MATTHIES, 1990). Après avoir insisté (ABRY et LALLOUACHE, 1991) sur les différences dues au contrôle de la jointure [≠] – qui apparaît inévitablement dans les groupes de consonnes complexes habituellement manipulés dans les expériences sur cette anticipation (dans notre cas [kstsk], [kssk], etc.) –, nous avancerons ici l'idée que la durée du mouvement est fortement expansible (allant dans notre cas jusqu'à tripler), mais relativement peu compressible. Le résultat phénoménologique de cette tendance est que le modèle LA tient ... tant que la suite de consonnes intervocaliques n'est pas trop longue, c.à.d. jusqu'à ce «point» où l'expansion n'est plus possible. Et c'est à ce «point» que le contrôle de la jointure peut changer complètement le profil du mouvement.

1. INTRODUCTION

Après avoir traité plusieurs milliers d'images (8600 pour cette seule étude) sur les lèvres d'un locuteur français sélectionné, enregistré à plusieurs mois d'intervalle, peut-on se faire une idée, autre que celles qui ont actuellement cours, sur les modèles d'anticipation? Sur un sujet controversé, où les auteurs ont expliqué tour à tour leurs différences de conception par les différences de leurs résultats, imputées à des variables indépendantes linguistiques et individuelles, nous nous contenterons de n'avoir à traiter aujourd'hui que de variables expérimentales (LUBKER et GAY, 1982).

PERKELL (1990) a testé systématiquement, pour l'anglais, les différentes versions des modèles dits *look-ahead*, *time-locked* et *hybrid*, avec une procédure renouvelée, utilisant les événements de vitesse et d'accélération.

Dans une séquence [iC_1...C_i...C_nu], voici le timing de ces événements que devraient permettre de prévoir les différents modèles.

• Le modèle *look-ahead* prédit que le début de la protrusion des lèvres (événement v=0+) commence aussitôt que la voyelle de trait contraire [i] se termine. C'est le modèle de HENKE (1966) choisi par BENGUÉREL et COWAN (1974) pour rendre compte de leurs données sur le français. Les démarrages du geste dans la voyelle [i] sont censés ne pas infirmer le modèle, dans la mesure où l'intégrité acoustique distinctive de celle-ci reste préservée.

• Le modèle *time-locked* prévoit au contraire que la protrusion démarre à peu près à date fixe par rapport au début acoustique de la voyelle [y], et ceci quelle que soit la longueur de la chaîne de consonnes. Ce modèle a été tout spécialement défendu, contre le premier, par l'équipe de Haskins (BELL-BERTI et HARRIS, 1981).

• Le modèle *hybrid* (PERKELL et CHIANG, 1986) prédit que la protrusion commence bien aussi tôt que dans le *look-ahead*; mais en pente plus ou moins douce, son accélération véritable (γmax) se situant plus tard, et à date fixe par rapport au début de la voyelle, comme c'est le cas du début de la protrusion dans le *time-locked* (d'où son nom d'*hybrid*).

Notons que le *time-locked* a été, dans sa version la plus récente, reformulé en modèle dit *frame* ou de *coproduction*, pour tenter d'expliquer la variation des profils de mouvement de type *hybrid* par une superposition de gestes de protrusion proprement consonantiques (pour [s, t, l,...]) sur le geste vocalique (BOYCE *et al.*, 1990, 1991). Les profils hybrides seraient ainsi dus à l'émergence de ces gestes consonantiques dans les conditions où les gestes sont supposés moins se chevaucher, c.à.d. lorsque le tempo ralentit, lorsque l'accentuation est plus forte ou lorsque (comme dans [iC_1...C_i...C_nu]) l'augmentation du nombre de consonnes intervocaliques produit un intervalle plus grand entre les voyelles.

La conclusion de PERKELL (1990) est la suivante: «In spite of the large amount of observed variability, there are relationships in the data which allow us to reject strong versions of all three models as means of accounting for the results of this particular experiments». Plus positivement (selon PERKELL et MATTHIES, 1990), mais seulement pour celles de leurs données sur l'anglais qui ne contiennent pas de [s] (pour lequel, en accord avec BOYCE *et al.*, 1990, ils admettent «some consonant-specific protrusion [*sic*] effects»), il y aurait trois contraintes en compétition : (1) terminer le mouvement de protrusion durant la partie voisée du [u]; (2) utiliser une durée de mouvement préférentielle [TL, Haskins]; (3) commencer le mouvement de protrusion quand les contraintes perceptives, qui exigent que le [i] précédent reste non-arrondi, sont relâchées [LA].

Dans une publication illustrative des données que nous présentons extensivement ci-dessous (ABRY et LALLOUACHE, 1991), nous rejoignions, pour le français, les conclusions de PERKELL (1990) sur le rejet des versions fortes des trois modèles et tombions d'accord sur une version lâche du point (1) : le maximum de protrusion se produit préférentiellement aux environs du début de la voyelle. La présente publication a pour objet de trouver une alternative à l'explication de la variabilité par la compétition de TL (2) et LA (3), à laquelle Perkell s'est finalement rangé. (Nous ne nous prononcerons pas, par contre, sur l'existence d'un geste de *protrusion* spécifique du [s] en français ...).

Enfin, nous insisterons à nouveau (ABRY et LALLOUACHE, 1991) sur les différences dues au contrôle de la *jointure*, puisque les expériences sur l'anticipation labiale utilisent couramment le paradigme de l'augmentation du nombre de consonnes intervocaliques, avec des groupes qui ne peuvent pas être toujours tautomorphémiques dans les langues étudiées.

2. CORPUS ET ENREGISTREMENTS

Pour étudier l'anticipation à travers les chaînes de consonnes relativement longues qui peuvent exister en français non méridional (comme dans plusieurs langues d'Europe), il est bon de comparer tous les cas de combinaison jusqu'aux plus chargés en éléments. Se rencontrent, dans ce français, des suites de cinq ou six consonnes, celui-ci possédant les conditions de réalisation de tels groupes consonantiques à la *jointure* des mots, dans des phrases comme "N'est-ce pas un directeur de société mixte scrupuleux?" Nous sommes partis d'une suite en miroir de cinq consonnes [kstsk] pour contrôler les mêmes influences consonantiques sur les deux contacts vocaliques entourant cette suite [i....y]. On peut ensuite effacer le segment central [t], puis l'un des deux [s], etc., toujours en gardant le contact immédiat de nos voyelles avec la même consonne, jusqu'à la jonction pure et simple des deux voyelles: [kstsk] -> [kssk] -> [ksk] -> [kk] -> [k] -> [-]. Il est en outre possible de contrôler, dans une certaine mesure, selon les contraintes phonotactiques propres au français, la position de la jointure [≠] : [kst≠sk] (seule possible) ; [ks≠sk] (*id.*) ; [k≠sk] et [ks≠k] ; [k≠k] (*id.*) ; [≠k] et [k≠] ; [≠]

Nous avons choisi pour phrase-mère "Ces deux Sixte sculptèrent" (soit, en substance : "Ces deux papes du nom de Sixte firent des sculptures"). Noter que le choix d'une séquence de voyelles [...e...ø...i...y...ɛ...], alternativement non arrondies et arrondies permet de contrôler les changements de sens des paramètres sensibles au geste d'arrondissement.

On notera que pour des raisons de taille du corpus nous n'avons pas doublé ces items par les contrôles correspondants [ikstski], [ykstsky], etc. Ceci peut être critiquable (GELFER *et al.*, 1989), si l'on pense que les consonnes intervocaliques pourraient produire – en dépit de leur réputation de neutralité aux lèvres – une certaine protrusion. Mais à notre connaissance, cet effet de protrusion consonantique est surtout observé sur le suivi en protrusion de la lèvre *inférieure* (*ibid.* et BOYCE *et al.*, 1991), pour laquelle il correspond essentiellement à une protrusion *induite* par l'élévation de la mandibule; mais pratiquement pas sur la lèvre supérieure, que nous avons finalement choisie pour tester les modèles (nous restons peu convaincus par les minimes variations sur lesquelles BOYCE *et al.*, 1990, se fondent pour arguer d'une protrusion *supérieure* propre aux consonnes [s, t, l, ...]).

Deux séances d'enregistrement de ce corpus (avec 12 répétitions en ordre aléatoire, dont nous avons pu retenir, en éliminant divers «ratés», de 5 à 10 exemplaires) ont été réalisées avec le même sujet (JLS; pour les conditions d'enregistrement vidéo et audio, cf. LALLOUACHE, 1991), à six mois d'intervalle (déc. 1989 et juin 1990). Seules les consignes et le nombre d'items ont varié d'une séance à l'autre.

• Dans la première séance, le sujet prononçait tout d'abord la partie nominale de sa phrase («Ces deux Sixte...»), *puis* l'ensemble de celle-ci («Ces deux Sixte sculptèrent»). Cette séance ne comportait que les séquences extrêmes [ikstsky] et [iky].

• Dans la seconde séance, par contre, la phrase était prononcée d'emblée complète, le sujet faisant l'économie de l'amorce de celle-ci. Cette séance comportait toutes les séquences choisies plus haut.

3. ÉTIQUETAGE DES SIGNAUX AUDIO ET VIDÉO

Après numérisation du signal audio à 16 kHz, nous avons visualisé en synchronisme les paramètres mesurés, grâce à notre éditeur EDILAB.

Nous avons pu déterminer plusieurs événements acoustiques (ABRY *et al.*, 1985). La fin de la voyelle [i] (événement VVT ou *Vocalic Voice Termination*, déterminé par la disparition de la structure formantique vocalique) et le début du [y] (événement VVO ou *Vocalic Voice Onset*, déterminé par l'apparition de cette structure formantique) sont deux événements qui permettent de définir un intervalle d'obstruence (IO), produit par les strictions des consonnes dans le conduit vocal, entre les deux voyelles (Fig. 1), sauf bien entendu pour les suites [iy]. Les débuts de [i] et fins de [y] étaient, eux, toujours repérables (VVO [i] et VVT [y]).

Enfin nous avons sélectionné un autre événement acoustique, précédant de peu VVO [y] : la détente de la dernière consonne [k] de la séquence (CFO ou *Consonantal Frication Onset*, correspondant au début du bruit de plosion).

Parmi les huit paramètres qui sont, dans l'état actuel de nos traitements (LALLOUACHE, 1991), les plus fiables dynamiquement, nous avons examiné pour cette étude : de face, l'écartement (A), la séparation (B) et l'aire intérolabiale (S) ; de profil, les protrusions des lèvres supérieure (P1) et inférieure (P2).

Etant donnée (cf. *supra*) l'importance des influences des consonnes intervocaliques sur P2, S et B (cette dernière, lorsqu'elle n'est pas trop grande, évoluant de conserve avec A), le timing de l'anticipation d'arrondissemnent sera déterminé de préférence sur les mouvements de la lèvre supérieure, qui semble dépendre le plus directement de la commande d'arrondissement (LALLOUACHE, 1991). En outre, P1 étant particulièrement peu bruitée dans notre détection (ce qui est remarquable étant donnée sa faible dynamique : 8 mm en moyenne chez ce locuteur), il semble que nous puissions lui faire confiance pour indiquer les événements de l'anticipation des lèvres (cf. *infra*).

Le traitement de la centaine d'images environ, que représente chaque séquence, prend sur notre poste *Visage-Parole* une vingtaine de minutes (recherche d'images sur le magnétoscope incluse), un temps de traitement que nous avons pu ramener récemment à 7 minutes, ce qui reste encore relativement long.

Les événements cinématiques, nécessaires pour tester les modèles d'anticipation selon les critères définis par PERKELL (1990), ont pu être obtenus en lissant nos données brutes de position à l'aide de fonctions splines cubiques. Nous avons détecté manuellement sur les dérivées première et seconde : le début du mouvement de protrusion (v=0+; que nous pouvons préférer appeler maximum de rétraction pour [i]; d'où le terme neutre choisi ci-après), le maximum de l'accélération, le maximum de la vitesse et le maximum de protrusion (ces événements sont indiqués respectivement par Min.Protr., Max.Acc., Max.Vit. et Max.Protr., Figs 1 et 4); et mesuré les valeurs de l'amplitude de la protrusion (Ampl.) et de la vitesse (V.Max.) à l'événement Max.Vit. Ces dernières mesures visent à caractériser le profil de vitesse du mouvement (NELSON, 1983).

4. RÉSULTATS

Ne seront donnés ici graphiquement que les résultats sur l'ensemble des séquences pour toutes les séances (avec quelques commentaires par classe phonétique ou séance, si nécessaire).

Les événements acoustiques et cinématiques repérés nous ont permis de mesurer différentes durées pouvant révéler l'organisation temporelle des phases de l'anticipation. Sur la figure 4 n'ont été représentées que les relations temporelles des événements cinématiques de protrusion, référencés par rapport à la fin de [i] (VVT), en les exprimant en valeur algébrique du pourcentage de la base temporelle que constitue l'intervalle d'obstruence IO (ne figure bien entendu pas [iy]).

L'adoption des conventions de PERKELL (1990), représentant les événements (référencés par rapport à la fin de [i] ou au début de [y]) en fonction de IO, a pour inconvénient de souffrir d'un artefact statistique "tout-partie" sur les corrélations (BENOÎT, 1986), puisque IO a une borne commune avec ces événements dans [i...y] (PERKELL et MATTHIES, 1990, ont tenu compte, au moins en partie, de cet artefact). L'adoption d'une représentation en pourcentage de la base (utilisée par BENOÎT et ABRY, 1986; formalisée par GENTNER, 1987) permet d'éviter cet inconvénient.

Première constatation, les Max de Protrusion apparaissent bien aux environs du début acoustique de la voyelle (la régression est tracée à titre seulement indicatif; R=-0.30 à la limite du seuil de signification, pour $p<0.01$). Cela signifie simplement que le pic de protrusion ne se produit pas bien avant le but acoustique de la voyelle, ce qui pourrait arriver, si le geste de protrusion atteignait très vite son maximum et/ou présentait un *overshoot*. Notons que la représentation en pourcentage de la durée d'obstruence consonantique permet de tester la relativité de tous les écarts, y compris les petits dans [iky].

Cette première coïncidence étant constatée, il nous est loisible d'examiner la durée du mouvement (de Min.Protr. à Max.Protr.) en fonction de l'intervalle IO (Fig. 2). La relation que nous trouvons infirme un modèle *time-locked*, puisque la durée croît significativement avec l'augmentation du nombre de consonnes. Remarquons que l'intercept de la droite de régression calculée, 124 ms, est très proche de la valeur moyenne observée pour les transitions [iy], 147 ms, la différence entre la régression précédente et celle qui intègre ces observations (Y = 0.39X +134 ms) n'étant pas significative. On notera que les données à IO élevé, soit [ikstsky] de 89 et 90, sont les plus dispersées

Cette régression (Y = 0.42X + 124 ms), devrait nous permettre, si nous prenons la coïncidence Max.Protr. = VVO[y] (Fig. 4) comme une donnée, de calculer avec une bonne approximation la relation observée (aussi en Fig. 4) pour le début du mouvement (Min.Protr.), soit Y= -12400/X + 58. Effectivement celle-ci est très proche de la régression hyperbolique observée Y = -10088/X + 44.6 (avec un coefficient de corrélation R= 0.93). Ce qui s'interprète à la fois comme un succès de la coïncidence posée et de la constante calculée (124 ms) pour la transition [iy].

Remarquer que les événements Max.Acc. et Max.Vit. accompagnent cette relation (excepté un cas, où ces deux événements se produisent près du début de [y], commenté in ABRY et LALLOUACHE, 1991, Fig. 2, n°2). Ce qui signifie que la lecture plus fine de ces événements cinématiques, proposée par Perkell pour tester aussi la possibilité d'un modèle *hybrid* (à date fixe pour le pic d'accélération), ne s'avère pas plus pertinente que celle du seul début de mouvement (les phases du geste que ces événements déterminent croissent elles aussi en fonction de l'augmentation de l'intervalle IO).

En résumé, *phénoménologiquement*, le mouvement de protrusion : (1) atteint son max. autour du début de la voyelle arrondie [y]; (2) commence de plus en plus tôt, par rapport à [y], en fonction de

l'augmentation du nombre de consonnes intervocaliques (entraînant dans son expansion ses événements de pic d'accélération et de vitesse); (3) peut se produire après [i] (à 20-40% de la durée de la chaîne de consonnes [kstsk]), ou dès le début de cette voyelle (à - 40-80% de la consonne suivante [k]). Les relations temporelles (1) et (2) permettent d'expliquer entièrement (3). Autrement dit, à la fois le fait que le modèle ne soit pas toujours *look-ahead* (cf. Fig. 1); *et* qu'il le soit dans la plupart des cas, ceux où l'expansion du mouvement (2) permet de «couvrir» l'intervalle consonantique; de même dans les cas où cet intervalle est trop court, la présence de la constante de transition [iy] (2) explique que l'anticipation de [y] pénètre profondément dans la voyelle [i].

5. DISCUSSION

Nous pensons ainsi pouvoir proposer une alternative à PERKELL et MATTHIES (1990) pour expliquer la variabilité globale des comportements de l'anticipation d'arrondissement, qui ne soit pas une explication mi *look-ahead* [LA], mi *time-locked* [TL] ni même *hybrid.*[H] L'apparence généralement LA de nos réalisations, comme leur apparence parfois contraire à ce modèle (sans être favorable pour autant aux deux autres, TL et H), repose en fait sur une même tendance : la durée du mouvement de protrusion de la lèvre supérieure peut connaître une expansion linéaire en fonction du nombre de consonnes, mais à partir de la constante d'exécution qu'on lui trouvera en l'absence de toute consonne intervocalique. Ce mouvement est, en d'autre termes, expansible sans être compressible très en deçà d'une constante [iy], du moins pour les conditions ici testées (nous savons en fait qu'il peut se réduire encore un peu en débit rapide, et nous n'oublions pas qu'en français l'évolution en *glide* vers [jy] reste possible)

Nous pensons que certains raffinements bien nécessaires pourront être apportés à ce modèle, en particulier en ce qui concerne *l'effet de l'appartenance syllabique* des consonnes. Ainsi dans [ikssky], on observe parfois des profils de vitesse qui tendent vers un plateau, et en quittant [y] le [l] qui le suit se fait, semble-t-il, «abriter» par la protrusion de la voyelle, ce qui donne une «épaule» au mouvement, retardant ainsi la rétraction de la lèvre vers [ɛ] (nous ne parlons bien entendu pas de protrusions *dues* à [s] ou [l], cf. *supra*).

De même les problèmes de *jointure* pourraient sembler nécessiter des ajustements minimes. Minimes, ils le sont certainement (WIOLAND, 1985) s'il faut essayer de distinguer [i≠ky] *vs.* [ik≠y], et même dirions-nous mettre en évidence l'évident, p.ex. que [ik≠sky] est significativement plus long en tous points que [iks≠ky], puisque [s] est le plus long à l'initiale de mot. L'apport des jointures est plutôt de suggérer une vision différente de la chaîne parlée, celle-ci ne pouvant être conçue comme une simple suite de segments. Ainsi nos chaînes de consonnes sont produites selon différents groupements des actions articulatoires nécessaires pour les réaliser (FUJIMURA *et al.*, 1991). Cette rencontre de groupes à la jonction des mots est un produit qui peut être plus ou moins instable, selon les facilités de *contrôle prosodique* qui sont disponibles. Ainsi la dispersion de nos données (p. ex. Fig. 2) va nettement croissant lorsque nous nous approchons des groupes les plus complexes [kstsk]. C'est pour ceux-ci que nous avons pu mettre en évidence des stratégies très différentes selon les séances : 89 offre les moyens, par la procédure d'amorçage de la phrase (cf. *supra*), de garder un Max. de Vitesse proportionnellement constant dans l'intervalle IO, ce qui n'est pas le cas de la séance de 90 (ABRY et LALLOUACHE, 1991).

6. PROPOSITIONS POUR UN MODÈLE DE VISAGE PARLANT

«If one does not consider the problem of coarticulation, the animation will not appear natural». Cette assertion, tirée de *Communication and coarticulation in facial expression* (PELACHAUD, 1991, p. 77) est aussitôt mise en pratique par son auteur, en adoptant le modèle LA, sur l'exemple même de BENGUÉREL et COWAN (1974), *sinistr'structur'*. Pour un problème qui reste généralement difficile en synthèse (comment joindre les mots?), et tout particulièrement en synthèse visuelle (comment joindre les lèvres entre les mots, à travers la jointure?), nous pouvons apporter maintenant deux propositions, dont l'efficacité perceptive pourrait être testée. Dans un modèle composite (ici même, WOODWARD *et al.*, 1992), une fois générée la durée de la chaîne de consonnes intervocaliques, l'anticipation d'arrondissement peut être calculée en utilisant la relation (Fig. 2) Durée du mouvement = f(IO). D'autre part, la forme du mouvement semble bien relever *globalement* d'un système dynamique du second ordre, vérifiant dans l'ensemble (avec la dispersion déjà relevée sur les exemples les plus longs) une constante c = (V.Max./Amplitude) x Durée du mouvement = 1.70 (Fig. 3). Rappelons que cette valeur reste proche de celle d'autres articulateurs (c=1.70 pour les mouvements laryngaux d'adduction/abduction, MUNHALL *et al.*, 1985; 1.80≤c≤1.90 pour les mouvements du dos de la langue, OSTRY et MUNHALL, 1985), étant idéalement π/2 (1.57) pour un 2nd ordre. Nous la comparons à celle des mouvements qui ont un profil de vitesse plutôt rectangulaire (1.00, comme celui de l'archet du violon) ou triangulaire (2.00). Toutes les hyperboles de la Fig. 3 sont calculées avec la même valeur asymptotique (0.80), constante dépendant de l'ajustement observé. Comme la constante de 124 ms dans [iy], celle-ci dépend sans doute du locuteur. En attendant de pouvoir traiter tout autant d'images sur d'autres lèvres, celui-ci pourra nous servir de *speaker de synthèse*, avec la possibilité de moduler l'Amplitude de sa protrusion suivant la relation dynamique observée.

Remerciements: à notre locuteur préféré, fidèle et patient, Jean-Luc Schwartz pour avoir accepté de «cuire» sous 1000 Watts halogènes et de se livrer, entre autres, au jeu subtil des jointures; à Joe Perkell, pour avoir aussi bien clarifié les prédictions des différents modèles.

RÉFÉRENCES

ABRY C., BENOIT C., BOË L.J. & SOCK R. (1985) Un choix d'événements pour l'organisation temporelle du signal de la parole, *4èmes JEP du GCP du GALF*, 133-137.

ABRY C. & LALLOUACHE M.T. (1991) Audibility and stability of articulatory movements. Deciphering two experiments on anticipatory rounding in French, *Proc. of the 12th Int. Congr. of Phonetic Sciences*, 1, 220-225.

BELL-BERTI F. & HARRIS K.S. (1981) A temporal model of speech production, *Phonetica*, 38, 9-20.

BENGUÉREL A.-P. & COWAN, H.A. (1974) Coarticulation of upper lip protrusion in French, *Phonetica, 30*, 41-55.

BENOIT C. (1986) Note on the use of correlations in speech timing, *JASA*, 80, 1846-1849.

BENOIT C. & ABRY C. (1986) Vowel-consonant timing across speakers, *Proc. of the 12th Int. Congr. of Acoustics*, A6-1.

BOYCE S.E., KRAKOW R.A. & BELL-BERTI F. (1991) Phonological underspecification and speech motor organization, *Haskins Laboratories SR-105/106*, 141-152.

BOYCE S.E., KRAKOW R.A., BELL-BERTI F. & GELFER C. E. (1990) Converging sources of evidence for dissecting articulatory movements into core gestures, *Journal of Phonetics*, 18, 173-188.

FUJIMURA O., ERIKSON D. & WILHELMS R. (1991) Prosodic effects on articulatory gestures. A model of temporal organization, *Proc. of the 12th Int. Congr. of Phonetic Sciences*, 2, 26-29.

GELFER C.E., BELL-BERTI F. & HARRIS K.S. (1989) Determining the extent of coarticulation : effects of experimental design, *JASA*, 86 (6), 2443-2445

GENTNER D.R. (1987) Timing of skilled motor performance: tests of the proportional duration model, *Psychological review*, 94, 255-276.

HENKE W. L. (1966) Dynamic articulatory model of speech production using computer simulation, Ph.D Thesis, MIT.

LALLOUACHE M.-T. (1991) *Un poste «Visage-Parole» couleur. Acquisition et traitement automatique des contours des lèvres*, Thèse de l'INP, Grenoble.

LUBKER J. & GAY T. (1982) Anticipatory labial coarticulation: Experimental, biological, and linguistic variables, *JASA*, 71 (2), 437-448.

MUNHALL K.G., OSTRY D.J. & PARUSH A. (1985) Characteristics of velocity profiles of speech movements, *Journal of Experimental Psychology : Human Perception and Performance*, 11, 457-474.

NELSON W.L. (1983) Physical principles for economy of skilled movements, *Biol. Cybern.*, 46, 135-147.

OSTRY D.J. & MUNHALL K.G. (1985) Control of rate and duration of speech movements, *JASA*,, 77 (2), 640-648.

PELACHAUD C. (1991) Communication and coarticulation in facial animation, Ph.D Thesis, University of Pennsylvania.

PERKELL, J.S. (1990) Testing theories of speech production : implications of some detailed analyses of variable articulatory data, in W.J. Hardcastle & A. Marchal (Eds), *Speech production and speech modelling*, Kluwer Academic Publishers, Dordrecht, Boston, London, 263-288.

PERKELL J.S. & CHIANG C. (1986) Preliminary support for a «hybrid model» of anticipatory coarticulation, *Proc. of the 12th Int. Congr. of Acoustics*, A3-6.

PERKELL J.S. & MATTHIES M.L. (1990) Timing of upper lip protrusion gestures for the vowel /u/, *JASA, Suppl. 1, 87*, S123 (Citations tirées de la révision. soumise au *JASA*, 13 nov. 1991).

WIOLAND F. (1985) Faits de jointure en français. Implication au niveaux articulatoire et acoustique. Incidences sur le plan des fonctions linguistiques, Thèse d'État, Strasbourg.

WOODWARD P., MOHAMADI T., BENOÎT C. & BAILLY G. (1992) Synthèse à partir du texte d'un visage parlant français (ce volume).

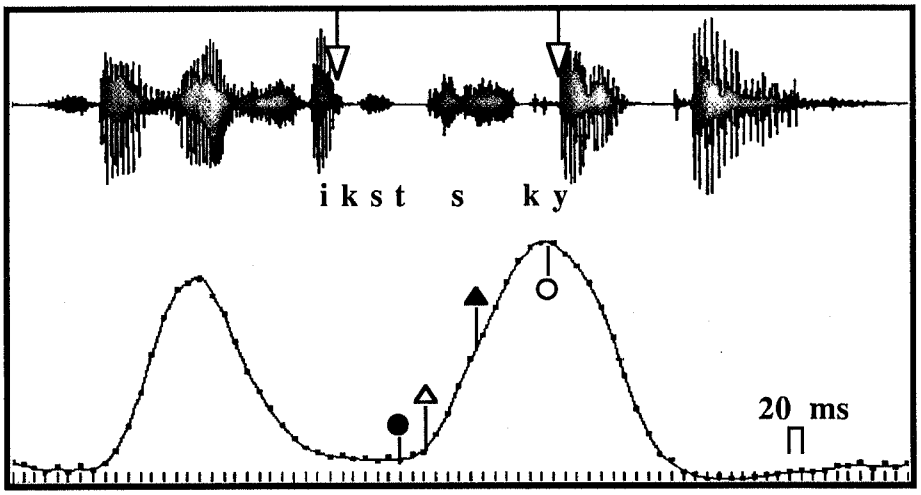
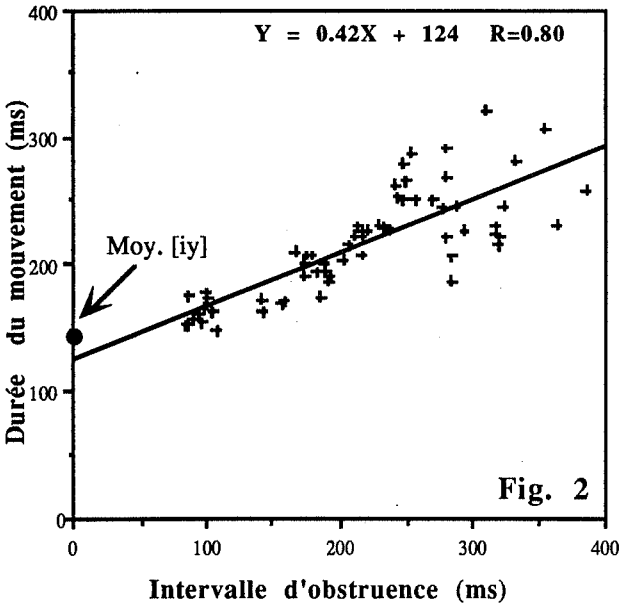
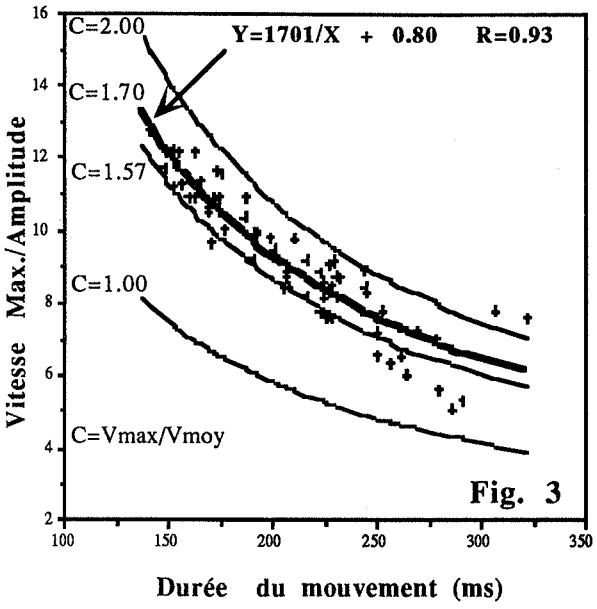
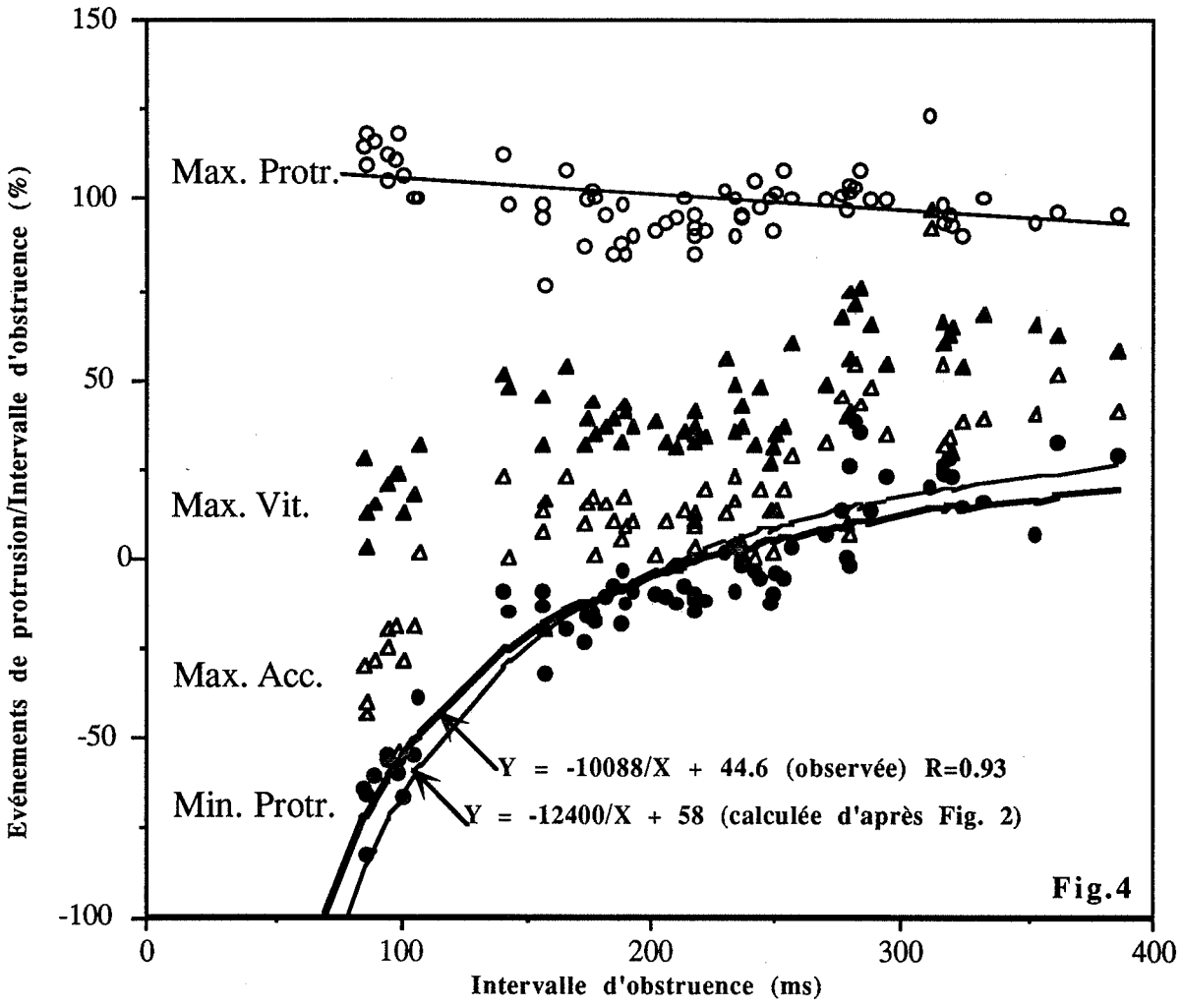


Fig.1 .— Exemple de signal de protrusion de la lèvre supérieure (1ère séance, déc. 1989), synchrone du signal audio à 16 kHz (les flèches indiquent l'intervalle d'obstruence). Les données brutes d'image (50 Hz) sont lissées par fonctions splines et les dérivées obtenues permettent de repérer les événements cinématiques (étiquetés par les mêmes symboles qu'en Fig. 4).







Pour les légendes des figures 2 à 4, cf. texte.

VERS DES PROTOTYPES ACOUSTIQUES ET ARTICULATOIRE DES 37 PHONÈMES VOCALIQUES D'UPSID

Nathalie VALLÉE & Louis-Jean BOË

INSTITUT DE LA COMMUNICATION PARLÉE
URA CNRS n° 368 INPG/ENSERG Université Stendhal BP 25 38 040 Grenoble Cedex 9 France

Résumé

Notre travail s'inscrit dans le cadre de la prédiction des systèmes vocaliques. Il a pour ambition de corroborer l'hypothèse selon laquelle la qualité des voyelles, pour un système vocalique donné, pourrait être directement liée à des capacités articulatoires et perceptives universelles.

Dans la continuité de notre recherche sur les critères de sélection des voyelles dans les langues naturelles, nous nous sommes fixés comme objectif d'associer des paramètres articulatoires aux valeurs acoustiques cibles des 37 phonèmes vocaliques universels (Maddieson, 1986), qui nous ont permis, dans une précédente étude (Vallée & al., 1991), de tester le modèle de prédiction de Schwartz & al. (1989). Cette étude présente 12 prototypes vocaliques qui couvrent l'espace maximal et à partir desquels il sera possible de proposer des standards pour l'ensemble des timbres de base d'UPSID.

INTRODUCTION

C'est l'hypothèse de l'efficacité fonctionnelle qui est traditionnellement et largement retenue pour expliquer la nature et le fonctionnement linguistique des systèmes phonétiques et phonologiques – dont il est facile de constater qu'ils ne sont pas dus au hasard. Cette hypothèse fait jouer à la distinctivité un rôle fondamental, chaque élément étant caractérisé par ses différences par rapport aux autres éléments du système. Dans cette approche, l'organisation systémique est donc privilégiée et la nature intrinsèque des éléments constituants minimisée.

En plaidant pour une analyse *substance-based*, Lindblom, dès 1972, proposait d'inverser les termes du débat en insistant sur l'importance des « bas niveaux » dans l'émergence des systèmes : « *we let the horse and cart change places* ». La sélection des unités phoniques se ferait sur la base de contraintes physiologiques, articulatoires et perceptives, qui permettraient d'expliquer, voire de prévoir, la structure phonologique des systèmes. Les différents modèles de prédiction proposés à partir de cette hypothèse (Liljencrants & Lindblom, 1972 ; Lindblom, 1986 ; Schwartz & al., 1989), ont surtout été améliorés par l'intégration de données psycho-acoustiques et par la prise en considération de critères de prégnance. Actuellement, il est tentant de faire appel à de nouvelles connaissances susceptibles d'améliorer ces prédictions.

Notre étude se situe dans le cadre d'une intégration de critères articulatoires. À terme, les résultats de cette recherche pourraient servir de données pour associer une dimension articulatoire aux critères acoustiques et perceptifs de distinctivité.

Il nous faut donc disposer de prototypes spécifiés à la fois au niveau articulatoire et au niveau acoustique. Compte tenu des données publiées dans la littérature, il est possible de disposer dans un espace 3-D des 37 timbres vocaliques qui permettent de décrire de façon symbolique l'ensemble des systèmes de la base UPSID (*UCLA Phonological Segment Inventory Database* ; Maddieson, 1986).

À partir d'une normalisation de l'espace acoustique vocalique, nous pouvons associer grâce à un modèle, les valeurs de paramètres articulatoires de commande correspondants. Celui de Maeda (1989) comporte 7 degrés de liberté : la séparation et la protrusion des lèvres, le corps, le dos et l'apex de la langue, la mâchoire et le larynx, il nous a semblé bien adapté à une description fonctionnelle. Pour ce faire, nous avons procédé, à partir d'un travail d'expertise, à l'établissement des prototypes vocaliques pour une douzaine de voyelles orales, bien répartis dans l'espace maximal. Cette étape a pu être menée à bien grâce aux *macro-variations*, fonctions d'interface entre entrée articulatoire et sortie acoustique qui permettent de

prévoir les conséquences acoustiques des gestes. Les macro-variations permettent de décrire les relations entre les formants et les articulateurs, ainsi que les paramètres géométriques cruciaux que sont le lieu et la dimension de la constriction, et l'aire aux lèvres (Boë & al., 1992). Elles vont permettre, à terme, à partir des prototypes déjà obtenus, de proposer l'ensemble nécessaire pour décrire les timbres de base d'UPSID.

1. PRÉDICTION DES SYSTÈMES VOCALIQUES

La justification des aspects universels des systèmes des sons a d'abord été recherchée dans l'étude des changements phonétiques. « *Les linguistes de la première partie du XIX^e siècle expliquaient l'existence et la régularité des changements phonétiques par une tendance générale de l'organisme vers un moindre effort* », à ce principe Roudet (1910, p. 342) ajoute « *la tendance à maintenir les distinctions phonétiques nécessaires à l'intelligence du langage* ».

Dans le même courant d'idées, Martinet (1955, p. 62) largement inspiré par les travaux de Groot (1931) en phonétique et phonologie, énonce ainsi le mécanisme général d'évolution des systèmes de sons : « *Les unités distinctives, les phonèmes qui coexistent, tendront naturellement à utiliser au mieux les latitudes que leur offrent les organes dits de parole ; ils tendront à être aussi distants de leurs voisins qu'ils est loisible pour eux de l'être tout en restant faciles à articuler et faciles à percevoir.* »

C'est sur ce même principe de discrimination des unités phoniques que s'appuie aussi la théorie du *contraste maximal* (Liljencrants & Lindblom 1972, p. 855) : « *The appeal of the principle of maximal contrast is no doubt based on the belief that vowels can serve as more efficient carriers of differences in meaning as they become more dissimilar, and the risk of confusing them decreases.* »

Cette théorie permet de générer par simulation, dans un espace de réalisation, un système donné : les voyelles s'organisent pour présenter un maximum de distinctivité globale, c'est à dire des distances maximales entre chaque paire de voyelles à l'intérieur du système. Le modèle de Liljencrants & Lindblom (1972) génère, sur ce principe, des systèmes théoriques proches des données typologiques de langues naturelles, mais seulement pour un nombre de voyelles inférieur ou égal à 6. Depuis il n'a cessé d'être complété et amélioré, en appliquant au principe de base, les résultats issus du développement des connaissances sur le système de perception humain. Les simulations ont progressé notamment grâce aux travaux de Lindblom (1975, 1986), Bladon & Lindblom (1981) et, au principe du contraste maximal, a été substitué celui de *contraste perceptif suffisant* (*sufficient perceptual contrast*) (Lindblom, 1986). Il semble mieux adapté au fait que les distances entre les voyelles, dans l'espace de réalisation, diminuent au fur et à mesure qu'augmente la taille des systèmes.

Grâce aux inventaires phonologiques qui portent sur de plus en plus de langues, des typologies telles que celle de Crothers (1978), Maddieson (1986) ou Vallée (1990), offrent des matériaux qui permettent de valider les modèles de prédiction. On constate alors que les facteurs psycho-acoustiques ne sont pas suffisant pour toutes les prédictions (Lindblom, 1986 ; Vallée & al., 1991), et qu'il faut rechercher dans le processus de production articulatoire, des critères susceptibles d'améliorer les simulations : paramétrisation de gestes, quantification du coût et de la complexité articulatoire (Lindblom & Lübker, 1985 ; Lindblom, 1986 ; Lindblom & Engstrand, 1990).

2. STABILITÉ ACOUSTIQUE DES SYSTÈMES VOCALIQUES

Il s'agit d'une notion qui a été développée dans le cadre de la recherche de nouveaux critères qui pourraient rendre plus efficaces les simulations des modèles de prédiction (Vallée & al., 1991). En accord avec les données issues de langues naturelles, Schwartz & al. (1989) proposent deux améliorations aux résultats obtenus jusque là par les modèles : éviter que trop de voyelles hautes ne se positionnent entre [i] et [u] et prédire une ou plusieurs voyelles antérieures arrondies, sans que le système ne s'équilibre avec une ou plusieurs voyelles postérieures ou centrales de même aperture.

Nous avons adopté l'hypothèse selon laquelle plus un système est fréquent dans l'inventaire des langues plus il est stable – stabilité appréciée au niveau acoustique par un critère d'énergie localement minimale (Schwartz & al., 1989 ; Vallée & al., 1991). A été testé avec le modèle de prédiction le comportement des systèmes vocaliques les plus fréquents à partir d'une typologie (Vallée, 1989), sur les données de la base UPSID de 317 langues (Maddieson, 1986). Ceci a été rendu possible grâce au modèle de prédiction dans l'espace tridimensionnel F1F2F3 de Schwartz & al. (1989). Ce modèle intègre le principe de dispersion de Liljencrants & Lindblom (1972) (c'est-à-dire la maximisation des distances intervoyelles comme fonction de l'énergie des systèmes), auquel a été ajouté un critère de focalisation. Le choix de ce critère repose sur l'étude de Schwartz (1987) qui tend à prouver que les voyelles qui présentent des F1 et F2 et/ou F2 et F3 et/ou F3 et F4 proches, sont préférées perceptivement (critère de prégnance). Pour que le modèle de prédiction ne soit pas falsifié il faut donc qu'il ne réorganise pas la structure d'un système fréquent, donc considéré stable.

Au sens des deux critères (maximum de distance et focalisation), 64% des systèmes fréquents se sont révélés acoustiquement stables, ce qui est un résultat résultat encourageant par rapport aux études

précédentes (Liljencrants & Lindblom, 1972 ; Lindblom, 1986 ; Lindblom & Engstrand, 1990). Cependant deux problèmes demeurent: celui des systèmes à 9 voyelles périphériques et celui des systèmes qui possèdent une voyelle centrale sans voyelle haute entre [i] et [u] (Vallée & al., 1991).
Comme cela a déjà été mis en évidence (Lindblom & Lübker, 1985 ; Lindblom, 1986 ; Lindblom & Engstrand, 1990), notre étude montre aussi que la prédiction reste limitée si les critères sont uniquement acoustiques et perceptifs. Il semble donc productif de spécifier des données articulatoires susceptibles de d'être intégrées dans des hypothèses explicatoires sur la structure des systèmes phoniques du langage.

3. PROTOTYPES ACOUSTIQUES

Le modèle de Schwartz & al. (1989) permet de tester la stabilité des systèmes vocaliques dans un espace multidimensionnel. Il faut donc disposer, pour chaque système, de la position de ses voyelles dans cet espace. La première étape a consisté à associer des valeurs formantiques (F1 à F3) aux 37 phonèmes vocaliques qui permettent de décrire la base UPSID.
Toute voyelle doit s'inscrire à l'intérieur de l'espace vocalique maximal. Celui-ci a été préalablement délimité à partir de résultats obtenus avec environ 60 000 simulations vocaliques, générées par le modèle de MAEDA. Les 37 voyelles ont pu être positionnées dans ce 3D à partir de l'expertise de données formantiques issues de diverses études acoustiques sur plus d'une trentaine de langues qui possèdent des systèmes vocaliques de différentes tailles (Vallée, 1990). Nous avons choisi de décrire chaque prototype acoustique par une hypersphère, standard pour toutes les voyelles, centrée autour de fréquences moyennes. La configuration obtenue respecte le « trou » observé dans les systèmes de langues naturelles autour de F1 = 300 Hz et F2 = 1000 Hz qui correspond à la zone formantique du conduit naso-pharyngal (Figure 1).

4. PROTOTYPES ARTICULATOIRES

Comme nous l'avons noté plus haut, les seules données acoustiques et perceptives se révèlent insuffisantes pour fournir une explication sur le contenu de certains inventaires phonologiques.
L'étude de la stabilité acoustique des systèmes vocaliques nous a conduit à établir les prototypes acoustiques des 37 voyelles utilisées par les phonologues pour décrire les systèmes vocaliques des langues du monde. Par une démarche identique à celle de Majid (1986), et grâce au modèle articulatoire de Maeda, il nous est possible d'inférer des données articulatoires à partir de cibles acoustiques. Nous nous sommes alors fixés comme but d'établir les prototypes articulatoires correspondant à chacune des cibles acoustiques.
Pour ce faire nous avons utilisé une méthode de simulation. Avec 7 degrés de liberté le modèle articulatoire de Maeda génère la coupe sagittale à partir de laquelle on calcule la fonction d'aires et la fonction de transfert du conduit vocal – laquelle fournit les paramètres acoustiques correspondants (formants). Les paramètres articulatoires de commande du modèle sont normalisés par rapport à une valeur moyenne (qui correspond au contour moyen du conduit vocal). Six d'entre-eux peuvent varier de +3 à -3 fois la valeur de l'écart-type par rapport à cette moyenne, le larynx ayant été limité entre $+1\sigma$ et -1σ.
Un déplacement des valeurs positives vers des valeurs négatives, correspond à un mouvement du haut vers le bas de l'articulateur, mis à part pour le corps de la langue, pour qui un tel déplacement correspond à un mouvement de l'arrière vers l'avant du conduit vocal. En fait, augmenter la valeur d'un paramètre tend à diminuer le degré de constriction.
Disposant des prototypes acoustiques, notre démarche a consisté à choisir des valeurs types pour chaque paramètre de commande de telle manière que chaque prototype soit correctement situé dans l'espace vocalique maximal, par rapport aux autres prototypes vocaliques.
L'évaluation de la fonction de transfert et des formants (Badin & Fant, 1984), nous permet donc de valider acoustiquement nos choix articulatoires. Par ailleurs, les voyelles peuvent être synthétisées (Feng, 1983 ; Castelli, 1988), ce qui permet une vérification auditive du timbre des prototypes retenus.
Pour établir des règles de commande, et donc pour choisir des valeurs de paramètres articulatoires, nous avons procédé à une large étude bibliographique qui nous a permis de disposer de données articulatoires. Bien souvent, ces études nous fournissent des données articulateur par articulateur et c'est dans une procédure d'ajustements itératifs, que l'on retrouve la cible acoustique prototypique.
Soulignons ici les principales difficultés auxquelles nous nous sommes heurtés. Elles sont dues essentiellement aux manques de données concernant :

- les sons peu fréquents dans les langues (tels que [ɵ], [ɜ], ...), qui sont peu ou mal connus ;
- les articulateurs les moins décrits pendant la phonation (exemple, l'estimation difficile de la position du larynx) ;
- la variabilité inter- et intra-locuteur pour une même langue ; la variabilité inter-linguistique (exemple, beaucoup de langues ont un [u] dont l'articulation est très avancée (Fischer-Jørgensen, 1985, p. 80).

Enfin il faut préciser que la relation entre formants et acoustique peut être localement non linéaire (Boë & al. 1992), et donc que les ajustements sont parfois délicats (exemple, le cas du dos et de l'apex pour certaines voyelles). La sensibilité d'un formant à une variation d'un paramètre dépend du degré de variation et du

paramètre modifié. C'est pourquoi nous avons établi, en premier lieu, les prototypes articulatoires de [i], [a], [u] et nous avons utilisé les fonctions de macro-sensibilités des trois premiers formants afin de récupérer la cible d'autres voyelles. En d'autres termes nous avons, pour chacune de ces voyelles, fait varier un à un les paramètres articulatoires de commande autour des cibles prototypiques, avec un pas de 0.25σ entre les valeurs maximales +3σ et -3σ (+1σ et -1σ dans le cas du larynx). Les trajectoires que l'on obtient (macro-variations), dans les plans F1F2, F2F3 et F3F4, autour de la position acoustique cible, et qui traduisent la sensibilité des formants aux modifications des valeurs de paramètres, nous permettent de déduire le déplacement d'un articulateur pour atteindre la cible acoustique d'autres voyelles (cf. Tableau 2).

5. APPLICATIONS

5.1. Les prototypes

La recherche de standards vocaliques correspondant aux 37 voyelles, fournit la matière première à l'établissement d'une « carte » articulatoire (au sens de figuration et repérage) comme on possède la « carte » acoustique (espace formantique), pré-requis pour l'extension des modèles de prédiction aux paramètres de commande.

La mise en évidence d'un classement hiérarchique des articulateurs pour toutes les voyelles (positions extrêmes et intermédiaires) repose le problème de la description traditionnelle des sons qui ne renvoie pas tout à fait à la réalité articulatoire. En effet, les traits articulatoires sont toujours un sujet de débat (Lass, 1984 ; Fischer-Jørgensen, 1985) qui reste ouvert et incontournable.

5.1. Les macro-variations

L'étude de ces fonctions d'interface entre l'articulatoire et l'acoustique permet de décrire les relations entre formants et articulateurs, ainsi que le lieu de constriction, l'aire de constriction et l'aire aux lèvres, qui sont de bons descripteurs géométriques (Boë & al. 1992).

Bénéficiant d'un quadrillage fin de l'espace acoustique grâce au nombre important de prototypes, les macro-variations autour de ces cibles (paramètre par paramètre et voyelle par voyelle) permettent une exploration fine de la relation articulatori-acoustique.

Par exemple, il est facile de repérer les paramètres qui ont le plus d'influence sur la sensibilité acoustique des voyelles : il s'agit du dos de la langue responsable du lieu d'articulation, ainsi que la mâchoire et la séparation labiale des lèvres qui déterminent la sortie du conduit vocal. En revanche, le mouvement de protrusion ou de rétraction des lèvres, et le mouvement de montée ou de descente du larynx, ne provoquent pas de grandes perturbations formantiques, que ce soit pour des voyelles arrondies ou non. La mise en évidence des paramètres les plus influents (c'est-à-dire ceux qui ont le plus d'amplitude sur la variation des valeurs formantiques) nous fait progresser vers une classification différenciée des voyelles dans l'espace acoustique.

L'étude des macro-variations fait ressortir les zones acoustiques plus ou moins sensibles à la variation d'un paramètre. Ainsi il est possible de localiser les zone de plus grande stabilité et les zones instables, et c'est, il faut le rappeler, un des critères central proposé par Stevens (1972 ; 1989) dans la *Théorie Quantique*.

CONCLUSION

Notre étude s'inscrit dans la recherche de nouveaux facteurs qui permettent d'améliorer les modèles de prédiction des systèmes vocaliques dont l'objectif est d'expliquer la formation et l'évolution des systèmes de sons dans les langues naturelles.

C'est un travail qui s'insère dans le cadre de la description et de l'explication des contraintes qui justifient la présence des voyelles universellement favorisées. La combinaison de mesures quantitatives de distinction perceptive et de complexité articulatoire nous semble productive dans le développement des modèles de prédiction.

Plus précisément notre étude peut permettre d'avancer dans le débat de l'explication des universaux, ainsi résumé par Lindblom & Engstrand (1990) : « *Are phonetic attributes selected because they are stable or because they are sufficiently different ?* »

Mais les deux critères sont-ils vraiment inconciliables?

RÉFÉRENCES

BADIN P. & FANT G. (1984), *Notes on Vocal tract Computations*, STL QPSR, 2-3, 53-108.

BLADON A. & LINDBLOM B. (1981), *Modeling and Judgements of Vowel Quality Differences*, J. Acoust. Soc. AM. 65, Vol 5, 1414-1422.

BOË L.J. PERRIER P. & MORRIS A. (1992), *Une prédiction de l'audibilité des gestes de la parole à partir d'une modélisation articulatoire*, XIX^{e} JEP.

BOË L.J. PERRIER P. & BAILLY G. (1992) *The Geometric Vocal Tract Variables Controlled for Vowel Production: Proposals for Constraining Acoustic-to-Articulatory Inversion*, Journal of Phonetics 20, 27-38.

CASTELLI E. (1988), *Caractérisation acoustique des voyelles nasales du français. Mesures, modélisation et simulation temporelle*, Thèse de Docteur Ingénieur, INP Grenoble.

CROTHERS J. (1978), *Typology and Universals of Vowel Systems in Phonology, In Universals of Human Language*, H.J. Greenberg (Ed), Standford University Press, Stanford.

FENG G. (1983), *Vers une synthèse par la méthode des pôles et zéros*, XIII^{e} JEP, 155-157.

FISCHER-JØRGENSEN E. (1985), *Some Basic Vowel Features, their Articulatory Correlates, and their Explanatory Power in Phonology, in Phonetic Linguistics*, 79-99, V.A. Fromkin (Ed), Academic Press, Orlando.

GROOT A.W. (1931), *Phonologie und phonetik als funktionswissenschaften*, TCLP 4, Prague.

LASS R. (1984), *Phonology: An Introduction to Basic Concepts*, 75-147, Cambridge University Press, Cambridge.
LILJENCRANTS J. & LINDBLOM B. (1972), *Numerical Simulation of Vowel Quality Systems: The Role of Perceptual Contrast.*, Language 48, 839-862.
LINDBLOM B.(1975), *Experiment in Sound Structure*, VIII[e] ICPhS, Leeds.
LINDBLOM B & LÜBKER J. (1985), *The Speech Homonculus and a Problem of Phonetic Linguistics, In Phonetic Linguistics*, 169-192, V.A. Fromkin (Ed), Academic Press, Orlando.
LINDBLOM B. (1986), *Phonetic Universals in Vowel Systems*, Experimental Phonology, Ohala J.J. (Ed), Academic Press, Orlando, Floridia, 13-44.
LINDBLOM B. (1990), *Models of Phonetic Variation and Selection*, Phon. Exp. Res. Inst. Ling. Univ. Stockholm 8, 21-39.
LINDBLOM B. & ENGSTRAND Q. (1990), *In What Sense is Speech Quantal ?*, Phon. Exp. Res. Inst. Ling. Univ. Stockholm 8, 1-20.
MADDIESON I. (1986), *Patterns of Sounds*, Cambridge University Press, 2ème Ed., Cambridge.
MAEDA S. (1989), *Compensatory Articulation During Speech : Evidence from the Analysis and Synthesis of Vocal-Tract Shapes using an Articulatory Model, In Speech production and Modelling*, 131-149, W.J. Hardcastle & A. Marchal, (Eds), Academic Publishers, Kluwer.
MAJID R. (1986), *Modélisation articulatoire du conduit vocal. Exploration et exploitation. Fonctions de macro-sensibilité paramétriques et voyelles du français*, Thèse de Docteur Ingénieur, INP Grenoble.
MARTINET A. (1955), Économie des changements phonétiques, Francke (Ed), 3ème Ed., Berne.
ROUDET L. (1910), Éléments de phonétique générale, H. Welter (Ed), Paris.
SCHWARTZ J.L. (1987), *A propos des notions de forme et de stabilité dans la perception des voyelles*, Bulletin du Laboratoire de la Communication Parlée, Vol. 1A,159-190.
SCHWARTZ J.L., BOË L.J., PERRIER P., GUÉRIN B. & ESCUDIER P. (1989), *Perceptual Contrast and Stability in Vowel Systems: A 3-D Simulation Study*, Eurospeech 89, Paris, Vol. 1/2, 63-66.
STEVENS K.N. (1972), *The Quantal Nature of Speech : Evidence from the Articulatory-Acoustic Data, In Human Communication : A Unified View*, 51-66, David Jr. & P.B. Denes (Eds), McGraw-Hill, New-York.
STEVENS K.N. (1989), *On the Quantal Nature of Speech*, Journal of phonetics, 17, 3-45.
VALLÉE N. (1990), *Structure et prédiction des systèmes vocaliques*, DEA, Université Stendhal Grenoble
VALLÉE N. BOË L.J. & SCHWARTZ J.L. (1990), *Systèmes vocaliques : typologies et tendances universelles*, XVIII[e] JEP, 32-36.
VALLÉE N. BOË L.J. & SCHWARTZ J.L (1991), *Tendances universelles et stabilité des systèmes vocaliques*, XII[e] ICPhS, 3, 142-145.

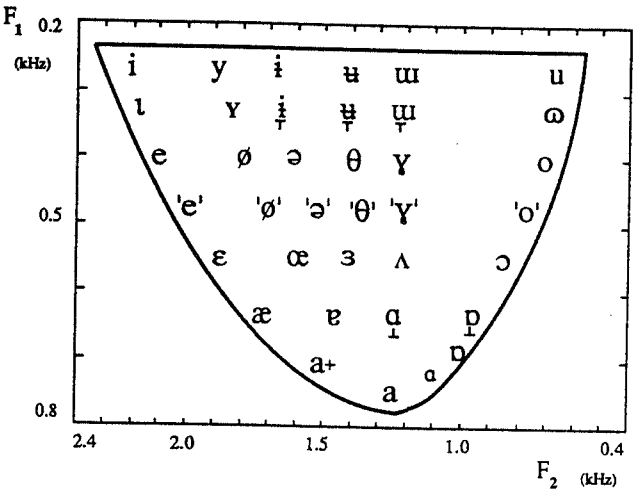


Figure 1 : Valeurs formantiques proposées pour les prototypes 37 voyelles d'UPSID (plan F1 F2).

	F1	F2	F3	F4
i	252	2191	3217	3840
e	409	1987	2781	3412
ɛ	563	1716	2396	3266
a	735	1210	2351	3492
u	265	787	1996	3110
o	387	773	2054	3263
ɔ	553	907	2096	3103
y	254	1849	2321	3116
ø	401	1609	2337	3415
œ	573	1434	2243	3371
æ	663	1658	2518	3555
ʌ	627	1302	2169	3221

Tableau 2 : Valeurs formantiques proposées pour les prototypes d'UPSID.

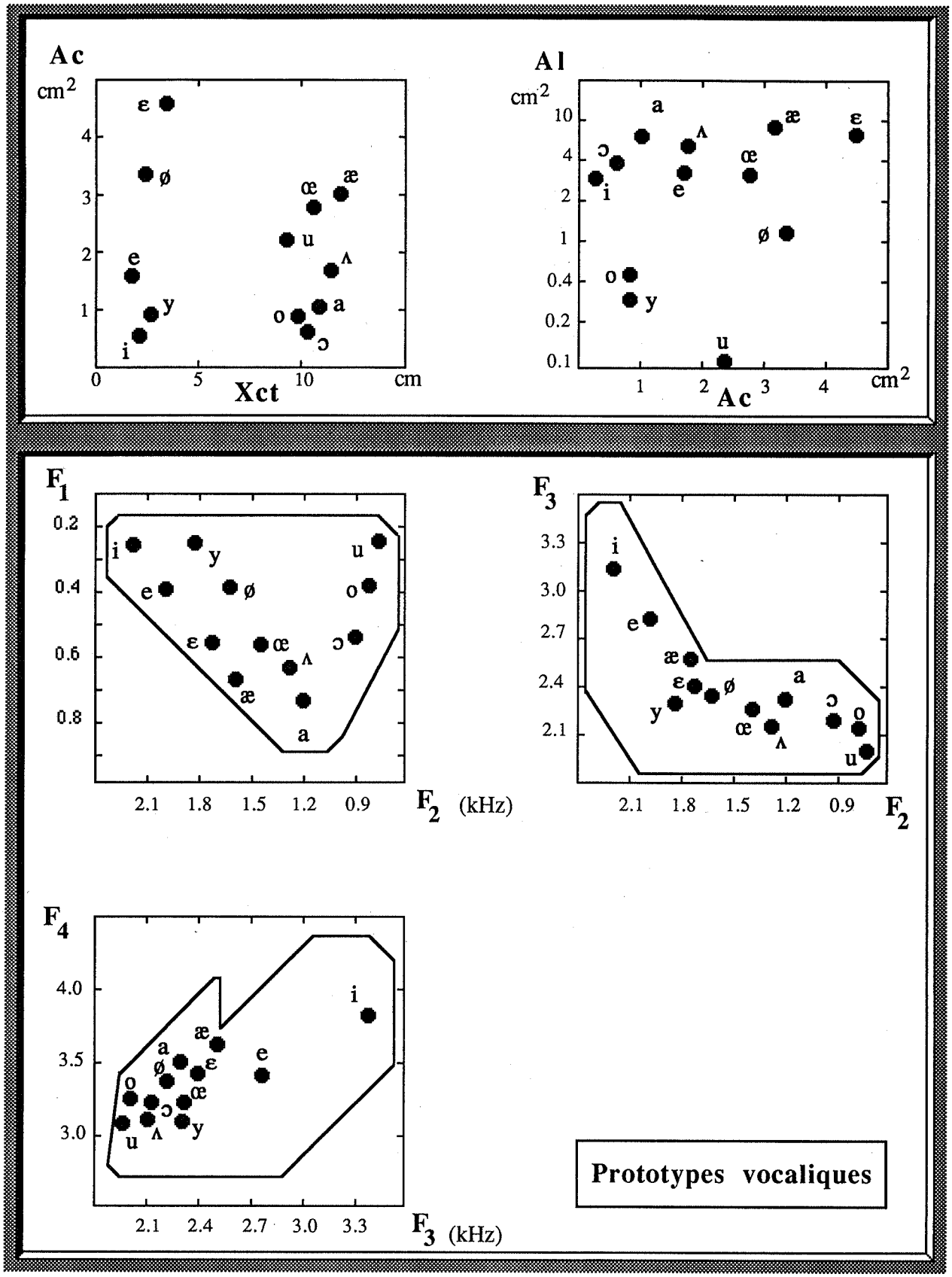


Figure 2. Propositions pour 12 prototypes vocaliques associables à UPSID, élaborés à partir du modèle de MAEDA à 7 paramètres. Les voyelles sont présentées dans les espaces :
– géométrique 3D Xc (position de la constriction par rapport aux dents), Ac (aire de la constriction), Al (aire aux lèvres),
– acoustique 4D F1 F2 F3 F4 ; les limites ont été relevées à partir d'une génération de 60 000 items.

19es J.E.P. - Bruxelles, 19 au 22 mai 1992.

CHUTE DE SEGMENTS ET TRACES DE DURÉE

CHANTALE TRÉPANIER ET DANIÈLE ARCHAMBAULT

UNIVERSITÉ DE MONTRÉAL

Résumé

En français québécois les voyelles hautes /iyu/ peuvent disparaître dans certains contextes. Des recherches ont démontré que différents facteurs peuvent jouer un rôle dans ce phénomène. Notre étude veut vérifier plus particulièrement le rôle de la vitesse de débit sur la chute des voyelles hautes en français québécois. Comme nous avons pu constater, grâce à des tests de perception, que la chute des voyelles hautes ne semble pas entraîner de confusion chez les auditeurs, nous avons chercher quels pourraient être les indices qui permettent aux auditeurs de bien reconnaître les mots malgré la chute de ces segments. Les analyses acoustiques que nous avons faites, nous ont permi de déterminer avec précision le nombre de chutes et de déceler différentes traces laissées par les voyelles lors de leur chute.

1. INTRODUCTION

Le français québécois présente un système vocalique élaboré dans lequel les oppositions de durée sont encore fonctionnelles. De ce fait, les voyelles présentes des patrons de durée variés et extrêmes. Ainsi, on peut retrouver d'une part, des voyelles excessivement longues telles que les voyelles nasales /ɛɑɔœ/ et les voyelles orales /oɑø / en syllabes fermées alors que d'autre part, certaines voyelles, telles que les voyelles /iyu/ ont tendance à s'élider. Ce dernier phénomène est suffisamment fréquent pour en faire une caractéristique importante du français québécois.

Plusieurs auteurs se sont penchés sur les phénomènes touchant les voyelles hautes; ceux-ci ont trait au timbre des voyelles, à leur désonorisation (Gendron, 1966; Santerre, 1975) mais c'est surtout leur tendance à s'élider qui a fait l'objet de nombreuses études (Dumas, 1978; Santerre, 1985). Certaines recherches ont aussi examiné l'effet de ces chutes sur la chaîne segmentale (Archambault, 1985; Santerre, 1975). Les différentes recherches ont confirmé l'ampleur du phénomène et ont donné un bon aperçu des modes et des contextes de disparition des voyelles (Hammond, 1980; Santerre, 1985; Cedergren et Lemieux, 1985).

La chute des voyelles hautes peut être reliée à plusieurs facteurs. Selon Santerre (1985), la disparition de ces voyelles peut être due aux caractéristiques même du système vocalique du français québécois qui en fait "un terrain propice à la chute des voyelles". Cedergren et Lemieux (1985), dans une étude sur la compréhension du changement linguistique, font une relation entre la désonorisation, l'abrègement et la syncope des voyelles hautes et associent ces processus au statut accentuel de la voyelle. Le phénomène est aussi plus fréquent dans certains contextes phonétiques (la cooccurrence avec les consonnes /t d v z/ par exemple).

La vitesse de débit a souvent été avancée comme un contexte propice à la chute des voyelles hautes. Gendron (1966) le remarque d'abord sur la désonorisation: "La fréquence des désonorisations chez les divers sujets est en rapport direct avec la rapidité du débit, au point qu'un de nos sujets, de prononciation pourtant excellente, mais qui a lu les mots très rapidement a dévoisé les voyelles beaucoup plus que ceux dont la lecture a été faite avec un débit plus lent." (p.52). Hammond a ausi étudié la chute des voyelles en rapport avec le débit et elle conclut que: "La probabilité de réduction augmente assurément avec le débit. Car plus le débit est rapide, plus l'escamotage de la voyelle est facile." (p.122).

Cependant, ce lien n'a jamais été examiné de façon systématique et semble plutôt être pris pour acquis étant donné que les réductions vocaliques sont d'une façon générale associées à un débit rapide. Toutefois Archambeault (1985) notait dans une étude sur les rencontres de consonnes identiques suite à la chute de segments, qu'un débit rapide n'assure pas nécessairement la disparition des voyelles.

La présente étude a donc pour but d'examiner de façon systématique le rôle de la vitesse de débit dans la disparition des voyelles hautes en français québécois. Nous voulons vérifier, à l'aide de variations imposées de débit, l'hypothèse selon laquelle un débit rapide s'accompagnerait d'un nombre accru de chutes de voyelles hautes. De plus, nous voulons examiner l'effet de ces chutes sur la structure phonétique des mots et le rôle que les traces possibles des voyelles disparues jouent dans la reconaissance de la forme phonologique. Nous aurons alors recours a des tests de perception.

2. CHUTE DES VOYELLES HAUTES ET VITESSE DE DÉBIT

2.1 Procédure expérimentale

Pour être en mesure d'évaluer plus particulièrement le rôle de la vitesse de débit, les voyelles à l'étude ont été insérées dans des contextes phonétiques propices à leur chute, soit en syllabe ouverte, précédées d'une constrictive sourde, un /s/, et suivies d'une occlusive sourde, /t/ ou /p/. Toutes ces voyelles sont en syllabe initiale de mot, en position inaccentuée.

Les voyelles à l'étude ont été insérées à l'intérieur de paires minimales mettant en opposition un mot avec la voyelle à l'étude et un mot sans cette voyelle (ex. "citation" /sitasjɔ̃/ ~ "station" /stasjɔ̃/). L'utilisation de paires minimales est utile pour assurer un contrôle des divers paramètres et pour retrouver les conditions nécessaires à cette recherche.

Chacun des huit mots (tableau I) a été inclus dans des phrases porteuses identiques, une fois en début de phrase et une fois en fin de phrase. Il y a ainsi quatre phrases par paire minimale. Le corpus compte donc seize phrases différentes.

Exemple:
Le **support** de Lucie est mauvais pour le dos.
Le **sport** de Lucie est mauvais pour le dos.
Vous devriez recommander plus de **support**.
Vous devriez recommander plus de **sport**.

La question principale de cette étude concernant l'influence de la vitesse de débit sur la chute des voyelles hautes, nous avons imposé aux locuteurs des variations dans les vitesses de débit. Nous leur avons demandé de lire les seize phrases du corpus à trois vitesses différentes: lente, normale et rapide. Chaque locuteur a donc prononcé un total de 48 phrases. Dix locuteurs ont participé à l'expérience, cinq hommes et cinq femmes, ce qui nous donne donc un corpus total à examiner de 480 phrases.

L'enregistrement s'est fait en chambre insonorisé sur un magnétophone Revox A77. Les locuteurs devaient d'abord se familiariser avec la liste des phrases et ensuite, en prenant leur débit naturel comme point de comparaison, ils devaient prononcer la même phrase à vitesse lente (le plus lentement possible sans faire de pauses) et à vitesse rapide.

Toutes les phrases du corpus ont fait l'objet d'analyses acoustiques sur un ordinateur Zénith AT à l'aide du programme Vis PC au laboratoire de phonétique de l'Université de Montréal. Ce programme permet entre autres de faire de la segmentation sur des spectrogrammes à 300Hz et d'enregistrer automatiquement la durée des segments à partir de cette segmentation manuelle.

2.2 Résultats

Chacun des dix locuteurs articulant ses variations de débit par comparaison à son débit naturel, les vitesses de débit intra et inter- locuteurs pour un même débit peuvent présenter des variations importantes. Ausi, avant d'examiner les taux de chutes, nous avons voulu vérifier d'abord si les locuteurs arrivent à produire les vitesses demandées et, ensuite, si ces vitesses sont asez semblables pour permettre des comparaisons. Nous avons donc calculé pour chacune des 480 phrases la vitesse de débit exprimée en syllabes par minute et la vitesse d'articulation (durée totale de la phrase moins les pauses). Le nombre de syllabes utilisé pour faire les calculs est le nombre de syllabes phonologiques car le nombre de syllabes réalisées phonétiquement est souvent difficile à déterminer, notamment à cause des chutes de segments.

$$\text{Vitesse de débit} = \frac{\text{Nombre de syllabes x 60}}{\text{Durée de la phrase (sec)}}$$

$$\text{Vitesse d'articulation} = \frac{\text{Nombre de syllabes}}{\text{Durée de la phrase - temps de pause}}$$

Étant donné la faible présence de pauses dans les phrases, la vitesse d'articulation seule s'avère une mesure satisfaisante pour l'analyse des résultats.

D'une façon générale les locuteurs arrivent à produire les vitesses en fonction des consignes données et celles-ci sont relativement uniformes d'une phrase à l'autre pour un même locuteur. En ce qui concerne les variations intra-locuteurs (tableau II), sauf pour quelques cas, les vitesses d'articulation sont bien distinctes et on remarque très peu de chevauchement, surtout en débit rapide. La vitesse d'articulation moyenne pour le débit normal est de 5,29 syl/sec (écarts: 4,13-6,12), de 4,3 syl/sec (écarts: 3,79-4,75) pour le débit lent et de 6,94 syl/sec (écarts: 5,86-7,36) pour le débit rapide.

Les analyses spectrales ont permis de déterminer de façon objective la chute ou le maintien des voyelles hautes pour chacune des phrases. Comme la caractéristique principale des voyelles au niveau acoustique est la présence des formants 1 et 2, c'est ce critère qui a servi à déterminer la chute ou le maintien des voyelles à l'étude.

L'examen du nombre de chutes pour le corpus entier fait ressortir un lien entre l'augmentation de la vitesse de débit et le nombre de chutes (tableau III). On remarque une augmentation des chutes avec l'augmentation de la vitesse de débit. Les pourcentages varient de plus de 15% entre le débit normal et le débit rapide, de 20% entre le lent et le normal et de près de 37% entre les débits lent et rapide. Le débit rapide est caractérisé par un taux de chutes excessivement élevé (toutes les voyelles à l'étude sauf deux sont tombées).

Donc un débit rapide favorise la chute des voyelles. Quant au débit lent, on ne peut affirmer, avec une incidence de chutes de plus de 60%, qu'il empêche la chute mais, si on le compare au débit normal, nous retrouvons une différence de 20%. C'est donc que le débit lent favoriserait plus que le débit normal ou rapide le maintien des voyelles hautes.

3. CHUTE DES VOYELLES HAUTES ET PERCEPTION DES MOTS

3.1 Procédure expérimentale

La chute des voyelles hautes modifie la chaîne segmentale du mot et pourrait donc entraîner des problèmes au niveau de la perception de ces mots. Nous avons donc eu recours à un test de perception auprès d'une quarantaine d'auditeurs.

Le test a été constitué de cent phrases choisies parmi les 480 du corpus. Dix phrases de chacun des locuteurs ont été sélectionnées. Le choix des phrases faisant partie du test de perception a été déterminé par le résultat des chutes. Donc, si un mot a un pourcentage de chutes plus élevé qu'un autre, ce mot apparaîtra plus souvent dans le test de perception.

Le test comprend 50 phrases porteuses d'un des mots à l'étude avec voyelle haute ainsi que les 50 phrases correspondantes sans voyelle haute. Sauf dans trois cas, nous n'avons gardé que des mots dans lesquels la voyelle haute était tombée. Les trois phrases où il n'y avait pas eu de chute de la voyelle haute ont été insérées à l'intérieur du test de perception comme point de comparaison.

Les cent phrases du test de perception ont été présentés dans un ordre aléatoire, cependant, les éléments d'une paire minimale ne peuvent se suivre non plus que le même mot placé au même endroit dans la phrase.

Quarante-et-un auditeurs, divisés en deux groupes ont passé le test de perception au laboratoire de phonétique de l'Université de Montréal. Les auditeurs sont toutes des femmes, étudiantes en orthophonie, âgées entre 20 et 30 ans et ne souffrant d'aucun problème auditif. Le test a été enregistré et retransmis sur un magnétophone Uher 2000. Le test a été effectué en salle calme, sans casque d'écoute.

Il s'agissait d'un test fermé, c'est-à-dire que pour chaque phrase entendue les auditeurs avaient le choix entre deux réponses seulement et ils devaient encercler celle qu'ils avaient choisie comme étant la phrase entendue. Le choix de phrases étaient toujours présentées dans le même ordre quelle que soit la phrase c'est-à-dire d'abord la phrase contenant l'élément de la paire minimale sans voyelle haute sous-jacente, ensuite la phrase avec le mot contenant une voyelle haute phonologique.

Exemple:
1.a) Dans ces églises on indique bien les **stations.**
b) Dans ces églises on indique bien les **citations.**

2.a) Vous devriez recommander plus de **sport.**
b) Vous devriez recommander plus de **support.**

Les tests de cinq des quarante-et-un auditeurs ont été rejetés parce que le français québécois n'était pas leur langue maternelle, un critère essentiel autant chez les auditeurs que chez les locuteurs. La compilation des résultats a donc été faite pour 36 auditeurs.

3.2 Résultats

Les résultats du test montre que de façon générale les phrases sont très bien reconnues et que, bien que la voyelle haute soit tombée, il n'y a pas de confusion chez les auditeurs. Sur les 100 phrases du test, 62 ont été bien reconnues par 100% des auditeurs et ce, quels que soient le débit, la place du mot dans la phrase et le type de voyelle.

Parmi les autres phrases, 28 ont un taux de reconnaissance variant de 72,2% à 97,2%. Ceci donne un total de 90 phrases sur 100 dont le taux de reconnaissance est supérieur à 64%, seuil au-dessus duquel les réponses des auditeurs peuvent être considérées comme n'étant pas l'effet du hasard avec une certitude de 5%.

Dans les dix phrases où il y a eu confusion, nous ne retrouvons pas seulement des phrases où le mot à l'étude a une voyelle haute sous-jacente tombée mais aussi des mots sans voyelle haute phonologique qui ont été entendus comme des mots ayant une voyelle haute. Ainsi, les mots "star" et "sport" par exemple ont été reconnus comme étant "cithare" et "support". En fait, sur les 47 phrases du test où la voyelle haute est tombée, 41 ont été parfaitement reconnues par les auditeurs. C'est donc plus de 87% de ces phrases qui sont bien reconnues. Pourtant, lorsqu'il y a chute de la voyelle haute, les deux membres des paires minimales sont semblables au niveau segmental. Ces résultats nous ont incité à croire que des indices de la présence sous-jacente de la voyelle pouvait se retrouver au niveau segmental, ce qui permettrait qu'il n'y ait pas de confusion.

3.3 Traces

Le test de perception a permis de constater que bien que la voyelle soit tombée, la perception des auditeurs n'est pas pour autant perturbée du moins dans la majorité des cas. Ce haut taux de reconnaissance porte à croire que des indices permettraient aux auditeurs de reconnaître les mots malgré la chute des voyelles.

En fait, lorsque la voyelle tombe, elle laisse parfois certaines traces sur la consonne qui la précède. Ces traces sont la labialité quand il s'agit de voyelles labiales comme le /y/ et le /u/ et la durée dans tous les cas.

Les voyelles labiales ont, entre autres, comme caractéristique de labialiser les consonnes qui les entourent. Ainsi, dans un mot comme "soupirante" (/supirat/), le /s/ initial sera labialisé par anticipation. Cette labialisation se manifeste au niveau acoustique par une baisse de la fréquence de la constriction du /s/. Par exemple, un /s/ dont la constriction est normalement en hautes fréquences, au dessus de 4 500Hz verra sa constriction s'étendre en moyennes ou en basses fréquences, vers 3 500Hz ou moins. L'examen acoustique des phrases montre que même dans le cas où la voyelle est tombée, la consonne est quand même labialisée. Ceci montre bien, d'une part que la voyelle est bien programmée bien que non réalisée et que, d'autre part, des traces de sa présence sont laissées sur la consonne.

Les figures 1 et 2 montrent les spectrogrammes de deux phrases où le mot à l'étude contient une voyelle haute à l'origine. Dans le premier la voyelle a été maintenue alors que dans le second elle est tombée. Dans les deux cas on peut voir très nettement que la fréquence du /s/ est très au-dessous de 4 000Hz. Cette labialisation de la consonne peut servir d'indice dans le cas de la chute des voyelles /y/ ou /u/ et permettre à l'auditeur de restaurer l'intégrité segmentale du mot.

Cependant, lorsque la voyelle disparue est un /i/, cet indice de labialité ne peut jouer puisqu'il s'agit d'une voyelle non labiale. Pourtant sur les 27 mots du test de perception avec voyelle /i/ sous-jacente, 23 ont été bien identifiés. On peut certainement supposer qu'un autre indice est présent dans le signal acoustique pour permettre aux auditeurs de retrouver le sens du mot malgré la chute de voyelles. Il est fort probable que cet indice soit la durée même de la consonne précédente.

Les analyses segmentales de durée montrent que le patron temporel diffère selon que la consonne initiale précède une voyelle haute phonologique, maintenue ou disparue au niveau phonétique. Les consonnes sont plus longues dans les mots avec voyelle haute sous-jacente (cithare /sitar/, soupirante /supirat/) que dans les mots sans voyelle phonologique (star /star/, spirante /spirat/. Cette diférence a trait à la structure syllabique des mots. Dans le premier cas nous retrouvons des syllabes de type CV (cithare /sitar/) alors que dans le dernier cas il s'agit d'une structure syllabique complexe de type CCV (star /star/). Le /s/ initial fait alors partie d'un groupe consonantique et sa durée s'abrège d'environ 30%, ce qui explique sa durée plus brève que dans l'autre membre de la paire minimale.

L'examen des durées segmentales montre que la durée des consonnes /s/ précédant les voyelles hautes à l'étude ne varie pas en fonction de la chute ou du maintien de la voyelle. En effet, qu'il y ait chute ou non, la durée (réelle ou relative) des consonnes et l'écart entre la durée de la consonne précédente et de la consonne suivante restent constants. Seule la durée du mot varie, puisque quand il y a chute, il y a un segment de moins. L'exemple suivant donne une illustration de ce phénomène pour le mot "soupirante".

Loc.7: /s/=17,86 /u/=7,14 /p/=9,82 Mot=41,4
Loc.6: /s/=18,5 /u/=0(chute) /p/=9,76 Mot=33,07

Pour le mot "spirante" nous retrouvons les durées suivantes:
Loc.6: /s/=13,01 /p/=12,38 Mot=34,83

Bien que les mots avec voyelle phonologique disparue et ceux sans voyelle haute sous-jacente présentent des structures segmentales identiques (cithare-[star], star-[star]), les mots se distinguent au niveau de la durée du /s/ initial qui est plus bref dans le dernier cas. La durée de la consonne /s/ initial dans le premier cas ne se comporte pas comme un membre d'un groupe consonantique. Cette durée plus longue constitue probablement un indice de la présence de la voyelle à un niveau sous-jacent.

5.CONCLUSION

Cette étude a permis de vérifier le lien entre un débit rapide et un nombre accru de disparition des voyelles hautes. L'analyse de la production de dix locuteurs montre qu'une accélération du débit entraine une augmentation dans la chute de ces segments et vice-versa. On note en moyenne une augmentation du nombre de chutes de l'ordre de 36% du débit lent au débit rapide. Il faut bien noter cependant qu'un débit lent n'assure pas le maintien de ces segments. On note en moyenne 60% de chutes en débit lent, ce qui montre bien que la vitesse de débit, si elle contribue à l'augmentation de chutes de segments, n'en est pas le principal facteur.

Parmi les 480 phrases recueillies, 100 ont été incluses dans un test de perception afin d'examiner si la chute de la voyelle dans le mot rend son identification difficile pour les locuteurs dans les cas où, suite à la chute de la voyelle le mot présente une structure segmentale identique à un autre mot de la langue (ex. cithare [star] -star [star]). Les résultats des tests montrent que malgré la chute, les auditeurs sont à même de reconnaître le mot dans 90% des cas. L'examen de la structure acoustique des mots reconnus montre que malgré la disparition de la voyelle, des traces de sa présence demeurent dans le mot. Ces traces sont la labialisation de la consonne précédente lorsque la voyelle disparue est un /y/ ou un /u/ et la durée plus longue de cette même consonne dans tous les cas. Cette durée plus longue de la consonne ne semble pas être le fait d'un allongement compensatoire. En effet, l'examen comparatif des durées des consonnes dans différents mots montre des durées semblables pour toutes les consonnes des mots avec voyelle haute sous-jacente, que la voyelle soit ou non tombée. En fait, c'est la consonne appartenant au mot sans voyelle haute sous-jacente qui s'abrège à cause de son appartenance à un groupe consonantique (ex. star - CCVC).

Le fait que la consonne précédente conserve sa pleine durée malgré la chute de la voyelle haute soulève de nouveau la question problématique de la resyllabification des éléments du mot. Ce comportement de la consonne laisse supposer qu'elle ne s'associe pas à la consonne suivante pour former un nouveau groupe consonantique. La syllabe initiale, malgré la chute de la voyelle serait donc maintenue avec la seule consonne comme support syllabique. Ces conclusions sont en accord avec celles rapportées par Santerre (1975) et Archambault (1985) dans leurs travaux sur les chutes de segments.

RÉFÉRENCES

ARCHAMBAULT Danièle, (1985) Production et perception des réductions de surface en français québécois, thèse de doctorat, Université de Montréal.

CEDERGREN Henrietta et Lemieux Monique, (1985) "La chute des voyelles hautes en français de Montréal "As-tu entendu la belle syncope?" in Les tendances dynamiques du français parlé à Montréal, tome 1, Gouvernement du Québec, Office de la langue française, Langues et Sociétés, pp.57-144.

DUMAS Denis, (1978) Phonologie des réductions vocaliques en français québécois, Thèse de doctorat, Université de Montréal.

GENDRON Jean-Denis, (1966) Tendances phonétiques du français parlé au Canada, Librairie C. Klincksieck, Paris,Les Presses de l'Université Laval, Québec, Québec, 254p.

HAMMOND Marie Andrée, (1980) La chute des voyelles hautes en français québécois, Mémoire de Maîtrise, Université de Montréal.

SANTERRE Laurent, (1975) "La disparition des voyelles hautes et la coloration consonantique en relation avec la syllabe en français québécois", Communication au 8ième Congrès International des Sciences Phonétiques, Leeds, Angleterre.

SANTERRE Laurent, (1985) "La chute des voyelles hautes en français de Montréal. Des voyelles inexistantes et pourtant bien perçues.", in Information Communication, vol.6, pp.5 à 21.

TABLEAUX

Voyelle	Paires minimales			
/i/	citation	station	/sitasjɔ̃/	/stasjɔ̃/
/i/	cithare	star	/sitar/	/star/
/y/	support	sport	/sypɔr/	/spɔr/
/u/	soupirante	spirante	/supirɑ̃t	/spirɑ̃t/

Tableau I: Corpus de paires minimales pour les voyelles hautes /iyu/

Locuteur	Vitesse d'articulation (syl/sec)		
	Lent	Normal	Rapide
1	3,79	4,95	7,35
2	4,34	5,89	7,13
3	4,44	4,13	6,3
4	4,75	5,81	7,01
5	3,96	4,5	6,13
6	4,34	5,59	7,35
7	4,42	5,59	7,32
8	4,71	6,12	7,56
9	4,09	4,66	5,86
10	4,19	5,61	7,36
Moyenne	4,3	5,29	6,94

Tableaux II: Vitesse d'articulation moyenne pour chacun des locuteurs (syl/sec).

Lent	Normal	Rapide	**Total**
49/80	65/80	78/80	192/240
61,25%	81,25%	97,5%	80%

Tableau III: Nombre de voyelles tombées par rapport au nombre d'occurrences.

FIGURES

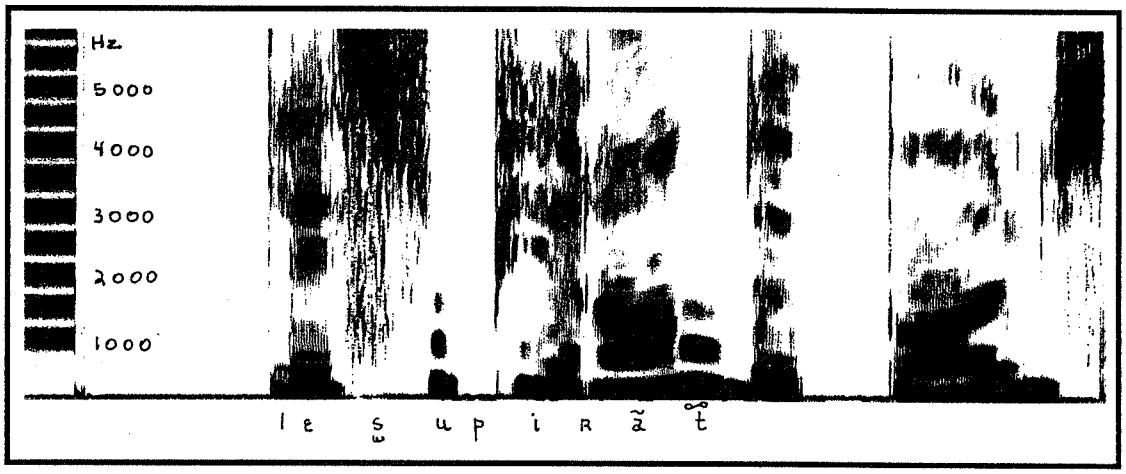


Figure 1: Spectrogramme du mot "soupirante" (filtre à 300Hz) où la voyelle /u/ n'est pas tombée.

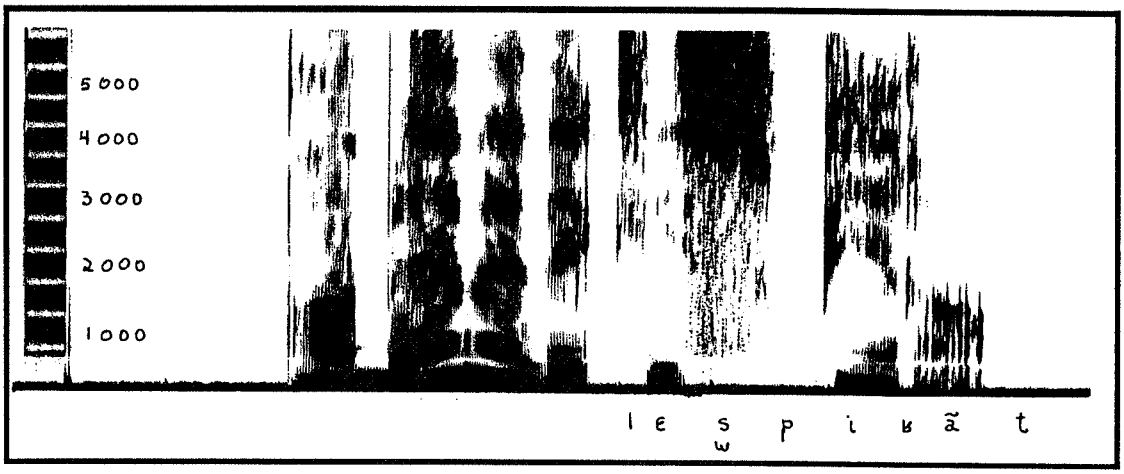


Figure 2: Spectrogramme du mot "soupirante" (filtre à 300Hz) où la voyelle /u/ est tombée.

SYNTHESE NON LINÉAIRE DE L'ONDE GLOTTIQUE

JEAN SCHOENTGEN*

INSTITUT DE PHONÉTIQUE
UNIVERSITÉ LIBRE DE BRUXELLES

* Fonds National de la Recherche Scientifique, Belgique

Résumé

Nous avons récemment proposé un modèle entrée-sortie du signal glottique. Le signal d'entrée est une cosinusoïde. Le modèle consiste en un couple de fonctions de distorsion non linéaires. L'objectif de l'exposé est de montrer comment réduire, à l'aide du modèle non linéaire, la quantité de données nécessaire pour décrire l'onde glottique d'un locuteur. Nous avons analysé et synthétisé les formes d'ondes obtenues par filtrage inverse glottique à partir d'un logatome produit par un locuteur masculin. Chaque cycle glottique a été caractérisé par son amplitude maximum, son facteur de forme et sa durée. Les indices de tous les cycles ont pu être copiés en utilisant les fonctions de distorsion de deux impulsions glottiques de référence et en modifiant les paramètres de contrôle du signal à l'entrée jusqu'à ce que la sortie du modèle présente les valeurs d'indices souhaitées.

Introduction

L'exposé traite de la synthèse du signal glottique à l'aide d'un modèle non linéaire. Le modèle est basé sur des techniques de représentation du signal (Schoentgen, 1990).

Les modèles linéaires tous-pôles ou pôles-zéros sont les plus souvent étudiés en rapport avec le signal de parole. Néanmoins, ils ne conviennent pas au traitement de l'impulsion glottique. Par conséquent, nous avons opté pour un modèle d'entrée-sortie non linéaire sans mémoire. En effet, la longueur et la largeur de la glotte sont petites par rapport à une longueur d'onde typique et les vibrations auto-excitées des cordes vocales ne peuvent être soutenues que parce que

(i) les forces aérodynamiques dépendent non linéairement de la position des cordes vocales,

(ii) les tissus des plis vocaux ont des propriétés non linéaires et

(iii) les cordes vocales entrent en collision.

Des oscillations auto-excitées peuvent exister dans des systèmes linéaires mais seules des propriétés non linéaires peuvent empêcher l'amplitude de croître jusqu'à ce que l'oscillateur soit détruit.

Notre modèle est un cas particulier d'une expansion en série de Volterra qui est une généralisation non linéaire d'un modèle linéaire à moyenne mobile:

$$y_t = \sum_{i=0}^{\infty} c_i x_{t-i} + \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} c_{ij} x_{t-i} x_{t-j} + \sum_{i=0}^{\infty}\sum_{j=0}^{\infty}\sum_{k=0}^{\infty} c_{ijk} x_{t-i} x_{t-j} x_{t-k} + \ldots\ldots \quad (1)$$

Dans (1), l'échantillon actuel y_t du signal dépend des échantillons passés du signal d'entrée stéréotypé x_t. L'hypothèse selon laquelle la glotte est un émetteur acoustique ponctuel suggère d'omettre

dans le modèle tous les échantillons du passé. Il en résulte une expression polynomiale (2) sans mémoire non linéaire en la fonction d'excitation x_t et linéaire en les coefficients c_i. L'expression porte le nom de formeur ou de fonction de distorsion.

$$y_t = \sum_{i=0}^{N} c_i \, x_t^i \, . \qquad (2)$$

Ce modèle permet d'analyser, d'encoder et de resynthétiser les impulsions glottiques avec une précision arbitraire pour autant que les coefficients c_i des formeurs soient connus. Nous avons montré que les c_i peuvent être calculés directement à condition que (Schoentgen, 1990):
(i) x_t soit une cosinusoïde
(ii) y_t soit périodique et soit ou paire ou impaire.
Pour cette raison, le modèle complet comprend deux parties, une pour la composante impaire et une pour la composante paire de y_t. Le formalisme décompose le signal en une excitation cosinusoïdale et un couple de formeurs qui dépendent seulement de la forme. Le signal original est récupéré lorsque la cosinusoïde est transformée par les formeurs (Figure 1). Environ 80 coefficients sont nécessaires afin de caractériser une impulsion glottique par un couple de formeurs. L'expansion de Volterra réduit, donc, peu les données qui concernent un cycle glottique. Par contre, il transforme les données spectrales pour leur donner une forme qui tient compte de la genèse non linéaire du signal et il ramène la synthèse à une consultation de deux tables.

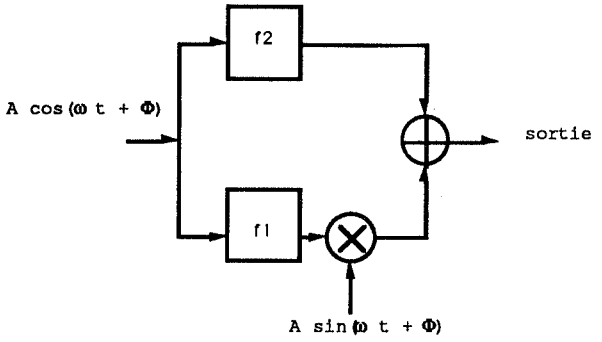


Figure 1

Dans cet exposé nous étudions comment réduire à l'aide du modèle de Volterra la quantité de données qui est nécessaire pour décrire l'onde glottique d'un locuteur. Nous montrons qu'il est possible de resynthétiser la totalité des cycles glottiques d'un locuteur à partir des modèles d'un petit nombre d'impulsions de référence. La procédure est légitime. En effet, les impulsions glottiques sont générées par un modèle non linéaire et les systèmes non linéaires ne sont pas caractérisés par une fonction de transfert. Cela veut dire qu'il n'existe pas de fonction unique qui décrit la réponse du système à n'importe quelle excitation. L'enveloppe spectrale cesse donc de jouer un rôle priviligié. On ne peut pas conclure que le modèle a été modifié en observant des changements dans la forme du signal. En d'autres termes, il est justifié de substituer le modèle de Volterra d'une impulsion glottique à celui d'une autre et d'expliquer les différences éventuelles entre les formes des deux impulsions par des modifications des paramètres de contrôle de la fonction excitatrice. Si cette substitution est possible, alors les donnéés concernant la forme du signal ont été bel et bien comprimées parce que moins de modèles que d'impulsions sont nécessaires. Les cycles peuvent être décrits par n'importe quel jeu d'indices qui convient. Dans cette étude, les indices étaient la période, l'amplitude et un facteur de forme.

Nous avons testé cette hypothèse en utilisant le modèle de Volterra pour synthétiser des cycles glottiques. L'onde glottique a été obtenue par filtrage inverse synchrone à partir du logatome [ama] produit par un locuteur masculin.

Ensuite, nous avons tenté de resynthétiser à partir de deux couples de formeurs de référence une séquence de cycles glottiques qui avaient les mêmes amplitudes, périodes et facteurs de forme que la séquence originale. Les couples de référence ont été choisis manuellement.

Modèle

Les effets d'un changements d'un paramètre de contrôle du signal excitateur - une cosinusoïde - sur le signal de sortie sont les suivants.
(i) La période du signal excitateur est égale à la période du signal de sortie.
(ii) Sur l'axe des fréquences, les harmoniques s'écartent lorsque la fréquence de la cosinusoïde excitatrice augmente et elles se rapprochent lorsque sa fréquence diminue. Les amplitudes des harmoniques ne sont pas affectées. Par conséquent, l'enveloppe spectrale évolue avec la période du signal.
(iii) Lorsqu'un couple de formeurs fixes est excité par une cosinusoïde dont l'amplitude A est différente de 1, le spectre du signal de sortie est plus pauvre que l'original en composantes fréquentielles si A<1 et plus riche si A>1.
(iv) L'évolution du spectre du signal de sortie avec l'amplitude A du signal d'entrée dépend de la phase ϕ du cycle. En effet, ϕ est un deuxième paramètre de contrôle (à côté de A) de la forme du signal. Le décalage temporel qui résulte d'un changement de ϕ est compensé en excitant le modèle par une cosinusoïde décalée de ϕ dans le sens opposé.

Les coefficients de Fourier d'un modèle décalé et non décalé sont reliés par deux expressions :

$$a_i(\phi) = a_i(0)\cos(i\phi) - b_i(0)\sin(i\phi)$$
$$b_i(\phi) = a_i(0)\sin(i\phi) + b_i(0)\cos(i\phi).$$

Les équations qui relient coefficients de Fourier et coefficients de Volterra sont les suivantes :

$$a_i = \sum_{j=1}^{N+1} m_{ij}\ c_j$$

$$b_i = \sum_{j=1}^{N} n_{ij}\ d_j\ .$$

Les a_i et les b_i sont les coefficients de Fourier et les c_j et les d_j sont les coefficients de Volterra. Les constantes m_{ij} sont égales aux coefficients d'un triangle de Pascal symétrique et les constantes n_{ij} sont égales aux coefficients d'un "pseudo-triangle" de Pascal antisymétrique. Le triangle antisymétrique est formé en calculant les différences entre deux colonnes adjacentes d'un triangle de Pascal normal. N est de l'ordre de 40.

Méthodes

L'objectif était de générer, avec un petit nombre de couples de formeurs, la forme d'onde obtenue par filtrage inverse synchrone du logatome [ama] produit par un locuteur masculin (Bailly, 1989). Les impulsions glottiques de référence ont été sélectionnées manuellement et leurs formeurs ont été calculés. Tous les autres cycles ont été caractérisés par leur amplitude, leur durée et leur facteur de forme. La définition du facteur de forme était la suivante :

$$ff = \frac{\sum_{i=1}^{40}\sqrt{a_i^2+b_i^2}}{\sqrt{a_1^2+b_1^2}} \quad .(3)$$

La modélisation et la synthèse de l'onde glottique ont été effectuées comme suit:
(i) Un cycle de référence a été sélectionné et ses formeurs f1 et f2 ont été calculés.
(ii) Les valeurs des indices de toutes les autres impulsions glottiques ont été reproduites en ajustant les paramètres de la cosinusoïde à l'entrée de f1 et f2. L'ajustement a été fait itérativement par un optimiseur jusqu'à ce que les valeurs des indices du signal de sortie du modèle soient suffisamment proches des indices des cycles d'origine. L'optimiseur était

l'algorithme polytope de Nelder et Mead (1965).
(iii) L'impulsion glottique a été synthétisée lorsque l'optimiseur avait convergé et lorsque les valeurs naturelles et synthétiques des indices étaient suffisamment proches.
(iv) Un nouveau cycle a été sélectionné parmi ceux qui n'avaient pas été correctement synthétisés à l'aide du modèle de référence précédent.

Les étapes (i) à (iv) ont été parcourues jusqu'à ce que toutes les impulsions aient été correctement synthétisées. Deux couples de formeurs ont été nécessaires lors de l'expérience (no 11 et no 19).

Finalement, les formes d'ondes ayant été synthétisées cycle par cycle, de faibles discontinuités apparaissaient aux endroits où des cycles contigus se rejoignaient. Ces discontinuités ont été supprimées par lissage au voisinage des points de contact.

Résultats

Les résultats de l'analyse sont indiqués par des carrés blancs sur les figures 2 à 4. Ils montrent la fréquence fondamentale, l'amplitude et le facteur de forme en fonction de leur numéro d'ordre dans la suite des périodes. Les limites entre segments (entre [a] et [m] et [m] et [a] respectivement) se situent approximativement aux périodes numéros 13 et 32. On peut voir que quelques valeurs sont irrégulières à l'attaque et lors des transitions entre segments. Il est possible que ces irrégularités soient la conséquence d'une analyse erronée par codage prédictif linéaire des transitions entre segments. Le codage par prédiction linéaire n'est en effet pas très apte à représenter des transitions.

Quoi qu'il en soit, nous avons décidé de reproduire les valeurs des indices acoustiques aussi précisément que possible avec un nombre de modèles de référence aussi petit que possible. Nous avons choisi deux impulsions (la 11^{e}, dans le [a] et la 19^{e}, dans le [m]), calculé les formeurs et déterminé les valeurs des paramètres de contrôle en minimisant la distance de Manhattan entre les indices analysés et synthétisés des impulsions 1 à 59. Les valeurs synthétiques des indices sont présentées sur les figures 2 à 4 par des losanges noirs. Elles sont superposées à leurs valeurs "naturelles" correspondantes.

En ce qui concerne la fréquence fondamentale, l'accord est parfait. La fréquence de la cosinusoïde excitatrice du modèle est ajustée exactement à la fréquence fondamentale de la sortie. La correspondance est bonne pour la vitesse volumique maximale et le facteur de forme. Le seul désaccord est observé à l'attaque et aux frontières entre segments. Le modèle reproduit difficilement le comportement erratique des valeurs naturelles à ces endroits. En fait, les valeurs synthétiques évoluent plus doucement. Une explication possible de cette évolution est que l'optimiseur était contraint à trouver des formes d'impulsions ayant des amplitudes proches aux instants ou deux cycles glottiques contigus se touchent.

Pour faciliter la comparaison, la figure 5 montre les formes d'ondes des périodes 39 à 45. Le signal du bas est le signal naturel obtenu par filtrage glottique inverse. Les deux autres trains d'impulsions ont été synthétisés avec les formeurs de l'impulsion n°11. Le signal du milieu a été synthétisé après une inversion par consultation de tables de la relation entre paramètres de contrôle et valeurs d'indices et celui du haut après inversion par optimisation de la même relation. Les signaux ont été lissés aux raccords des impulsions.

Discussion

Dans cette étude, nous avons sélectionné manuellement les cycles de référence. Toutes les autres impulsions ont été générées avec les formeurs de ces cycles. Bien sûr, dans un système opérationnel, il faudrait trouver automatiquement les modèles de référence et leur nombre. Une solution possible à ce problème consisterait à calculer les modèles de toutes les impulsions glottiques figurant dans un corpus d'apprentissage produit par un locuteur. Les modèles nécessaires pour synthétiser les impulsions glottiques

d'un locuteur pourraient alors être déterminés par quantification vectorielle.

Le nombre de modèles final dépend de la précision avec laquelle les formes d'ondes doivent être synthétisées. Il est clair que si la moindre différence entre deux impulsions ne peut pas être négligée, presque autant de modèles que d'impulsions seront requis. Par contre, pour un ensemble d'indices comme celui qui a été utilisé ici et qui constitue probablement le plus petit ensemble que l'on puisse envisager, les résultats indiquent que le nombre final de modèles peut être petit.

La quantification vectorielle nécessite que l'on puisse comparer deux impulsions et décider si elles ont été générées avec le même modèle. Dans le cas de signaux produits par un système non linéaire, nous avons montré qu'il est vain de baser cette décision sur leurs formes ou leurs spectres. La comparaison peut en revanche être basée sur les coefficients de Volterra. Mais, sachant qu'il existe une relation biunivoque entre les coefficients de Fourier et de Volterra, il est inutile de calculer une distance classique. Une solution possible à ce problème est la suivante. Supposons que l'on soit en présence de deux signaux dont les amplitudes et les spectres sont différents mais qui ont été générés avec le même couple de formeurs. Soit (c_n, d_n) et (c_n', d_n') les coefficients de Volterra correspondants. De plus, supposons que la phase n'a pas changé d'un signal à l'autre. Les différences entre les coefficients doivent alors provenir de différences dans les amplitudes des fonctions excitatrices. En d'autres mots, on a:

$$c_n' = c_n A^n$$
$$d_n' = d_n A^{n+1},$$

ce qui donne pour tout n :

$$A = \sqrt[n]{\frac{c_n'}{c_n}} \quad \text{et}$$

$$A = \sqrt[n+1]{\frac{d_n'}{d_n}} \quad (4)$$

A condition de prendre des précautions concernant le gain et la phase, on peut calculer (4) pour tout couple de cycles y_t et y_t' obtenus expérimentalement. Evidemment, il n'est alors pas garanti que les A soient les mêmes pour tout n. Nous pensons que (5) peut être une mesure servant à prédire si la forme d'onde y_t peut être générée avec le modèle de la forme d'onde y_t' (n et m sont fixes et n est différent de m).

$$\left| \sqrt[n]{\frac{c_n'}{c_n}} - \sqrt[m]{\frac{c_m'}{c_m}} \right| \quad (5)$$

Les invariants tels que (4) sont propres aux modèles de Volterra. Ils sont absents des modèles de formes d'ondes concaténées.

Remerciements

Une partie du présent travail a été menée à bien dans le cadre du projet MULTIDIF financé par l'ACCT (réf. 422/SG/C5).

Références

G. Bailly (1989), Communication personnelle

J. A. Nelder, R. Mead (1965), "A simplex method for function minimization", Computational J., 7, pp 308-313

M. B. Priestley (1981), "Spectral analysis and time series", Academic Press, London, p 869

J. Schoentgen (1990), "Nonlinear signal representation and its application to the modelling of the glottal waveform", Speech Communication, 9, pp 189-201

Figure 4

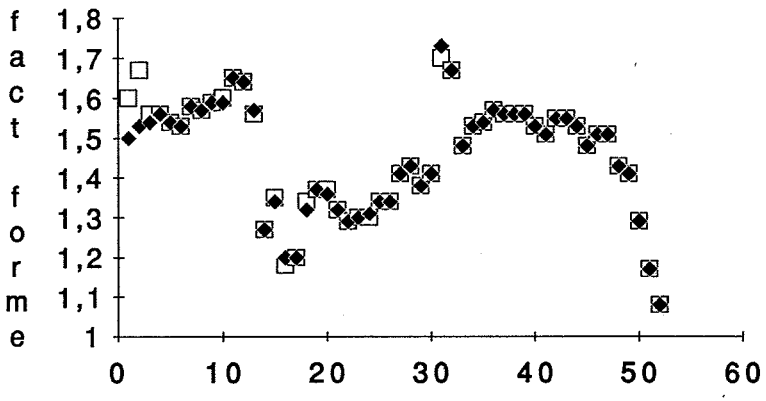
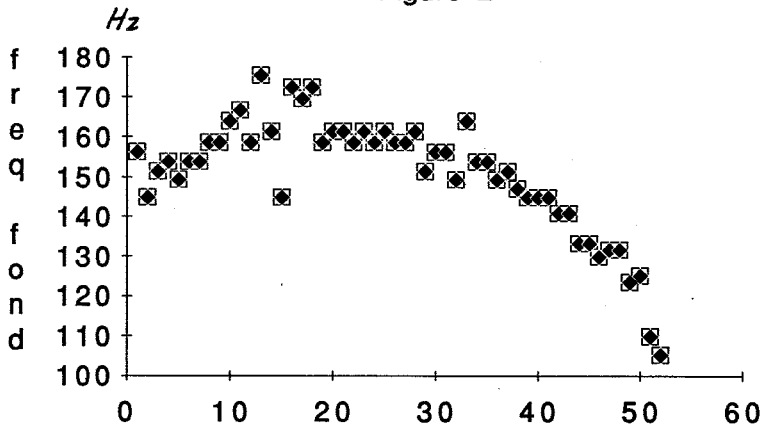


Figure 2



70

Figure 5

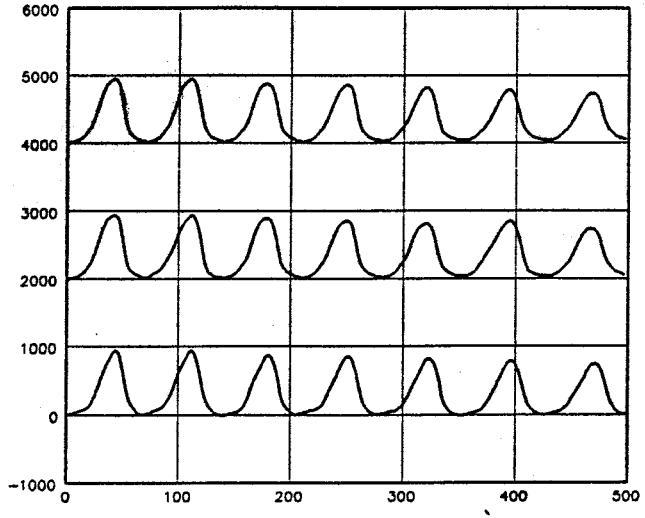
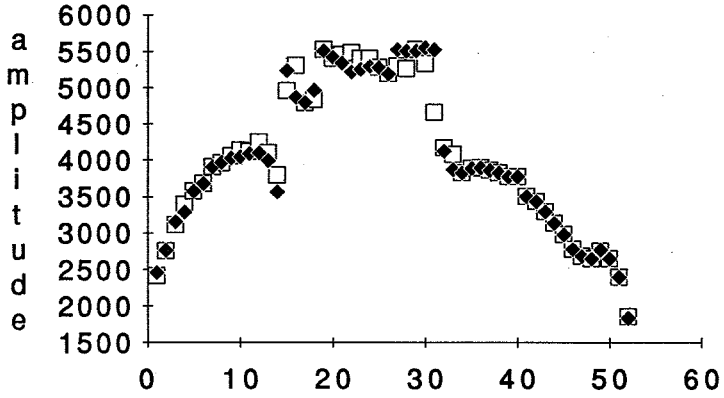


Figure 3



19es J.E.P. - Bruxelles, 19 au 22 mai 1992.

SYNTHESE DE BRUITS PAR FORMES D'ONDES FORMANTIQUES ALEATOIRES

G. Richard C. d'Alessandro S. Grau

LIMSI-CNRS, BP 133, 91403 Orsay Cédex

Résumé

Nous présentons, dans cette article, une nouvelle méthode pour la synthèse de la partie aléatoire du signal de parole et plus spécifiquement, les bruits de frication et les bruits de souffle. Cette méthode s'appuie sur les propriétés statistiques spectro-temporelles de ces bruits. On utilise, par analogie au bruit de Grenaille, des points de Poisson comme instants de déclenchement aléatoires de Formes d'Ondes Formantiques (FOF) dont les paramètres sont déterministes. Le caractère aléatoire n'est alors porté que par les instants d'occurence des FOF. Le bruit obtenu possède néanmoins les mêmes propriétés statistiques que le bruit blanc gaussien filtré. On donne des résultats expérimentaux sur les fricatives non-voisées et les bruits de souffle.

Mots clés: *Synthèse, Fricatives, Bruit, FOF.*

1 Introduction

Cet article présente une nouvelle méthode pour la synthèse de la composante aléatoire du signal de parole: bruit fricatif, bruit de souffle. Le propos est de représenter le bruit par les statistiques de ces maxima locaux d'énergie dans le domaine spectro-temporel. A la synthèse, le bruit est généré en sommant des ondelettes déterministes, dont seules les positions spectro-temporelles sont aléatoires, en accord avec les statistiques observées sur du bruit naturel. Il ne s'agit donc pas d'un modèle physique de la production du bruit, mais d'une représentation du signal produit, caractérisé par les indices acoustiques descriptifs qui semblent perceptivement pertinents. Pour les sons bruités continus considérés ici, ces indices sont:

1. les régions spectrales dominantes et leurs évolutions, qui représentent les maxima de la fonction de transfert de cavités acoustiques. Ces régions évoluent assez lentement et de façon déterministe.

2. la densité des maxima de l'enveloppe (aléatoire) du signal dans chaque région fréquentielle dominante. Cette densité evolue lentement.

La première partie rappelle les propriétés statistiques du premier et second ordre des bruits de la parole, restreints au bruit fricatif et au bruit de souffle. La décomposition du signal comme une somme d'ondelettes aléatoires se rapportant aux maxima spectro-temporels du signal est alors introduite. La seconde partie présente la nouvelle méthode de synthèse par ondelettes aléatoires. Les ondelettes sont choisies ici comme des Formes d'Ondes Formantiques (FOF). Après un rappel de la définition des FOF, on montre, en utilisant le formalisme du bruit de grenaille, que la génération de FOF suivant une loi de Poisson permet de synthétiser un bruit dont les propriétés statistiques sont identiques à celles du bruit blanc gaussien filtré. On montre également, que le caractère aléatoire est exclusivement porté par les instants d'occurence des FOF .

On discute ensuite de la justification acoustique de ce procédé de synthèse en montrant que la densité de maxima observée dans l'enveloppe d'un signal naturel est liée à la densité des FOF utilisée en synthèse. La troisième partie présente des résultats expérimentaux sur les propriétés statistiques des bruits de synthèse obtenus et sur la synthèse des fricatives non-voisées du français. La dernière partie conclue, en indiquant les évolutions possibles de la méthode présentée, en particulier son application aux bruits non-gaussien et non quasi-stationnaires.

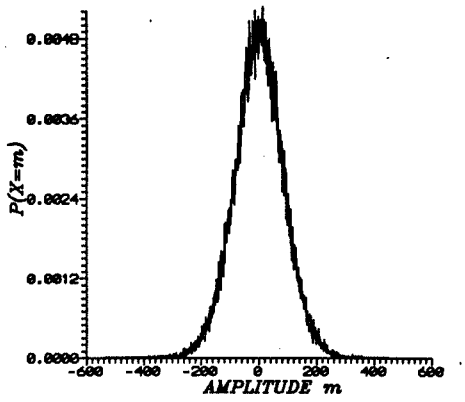


Figure 1: densité de probabilité de [f] tenu en contexte [ø]

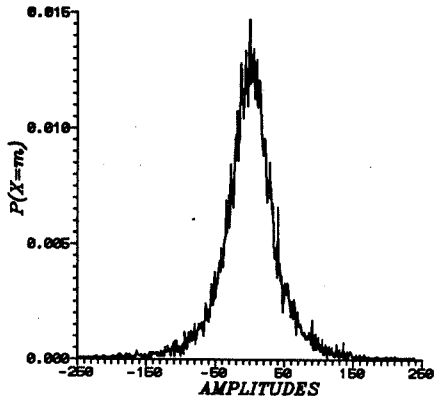


Figure 2: densité de probabilité d'un bruit de souffle

2 Proriétés statistiques des bruits en parole

Les trois grandes classes de bruit en parole sont:

1. les bruits transitoires: relâchement de plosives, bruit de bouche, coup de glotte.
2. les bruits continus: voix chuchotée, bruit de souffle, bruit d'aspiration et bruit de frication. Les bruits continus sont quasi-stationnaires puisque leurs propriétés statistiques évoluent relativement lentement dans le temps.
3. les autres bruits, qui possèdent une structure temporelle: voix grinçante (creaky voice), consonnes vibrantes ou fricatives voisées.

Seule la seconde catégorie sera considérée ici.

2.1 Statistiques du premier ordre

On suppose que les bruits étudiés peuvent être représentés par des processus stochastiques ergodiques. Cette hypothèse, dont la justification physique est hors de notre propos, permet d'obtenir des résultats statistiques significatifs pour la représentation du signal, en ne considérant que des moyennes temporelles. Les propriétés statistiques du premier ordre (densité de probabilité d'amplitude) et du second ordre (spectre de puissance) des bruits continus vont être rappelées.

Les densités de probabilité empiriques $p(x)$ sont obtenues par l'histogramme des amplitudes d'une réalisation assez longue du signal $x(t)$. Pour des bruits fricatifs ([f] , [s] et [ʃ]), ou des bruits de souffle, on obtient des densités gaussiennes de moyenne nulle et d'écart type σ_x:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma_x} \exp\left(-\frac{x^2}{2\sigma_x^2}\right) \qquad (1)$$

Cette distribution gaussienne reflète le processus de production du bruit, dû à un écoulement turbulent au voisinage d'une constriction dans le conduit vocal. Il faut noter que la densité gaussienne est une propriété particulière de ce type de bruit, car la densité de probabilité de la parole continue est tout à fait différente [5]. Les figures 1 & 2 montrent la densité de probabilité empirique d'un bruit fricatif tenu en contexte vocalique [ø], et d'un bruit de souffle obtenu à partir d'une voix feminine par décomposition harmonique/aléatoire (voir dans les mêmes actes [3]).

Aucune différence significative n'apparait sur les statistiques empiriques du premier ordre entre les différentes fricatives ou le bruit de souffle, si ce n'est le niveau du signal qui est représenté par σ_x.

2.2 Statistiques du second ordre: décomposition formantique

De nombreux travaux ont portés sur l'étude de la densité spectrale de puissance des bruits en parole (voir [11] pour une revue). Cette densité spectrale de puissance est imposée par la configuration déterministe du conduit vocal, puisque les cavités acoustiques, en avant et en arrière, de la constriction filtrent la source de bruit produite à la constriction. Selon la théorie acoustique source/filtre de production de la parole, la source d'excitation donne au signal son caractère aléatoire et gaussien, et le filtrage impose son spectre de puissance, que l'on peut décomposer suivant plusieurs régions dominantes associées aux formants.

On peut donc représenter le spectre de puissance du bruit grâce à la fonction de transfert, comportant des pôles et des zéros, du conduit vocal. En synthèse de parole, le synthétiseur à formants en parallèle est généralement adopté pour les sons non-voisés. On modélise, alors, le conduit vocal par un filtre digital de fonction de transfert:

$$H(z) = \sum_{i=1}^{M} \frac{G_i}{(1 - z_i z^{-1})(1 - z_i^* z^{-1})} \qquad (2)$$

avec M formants représentés par les paires de pôles conjugués (z_i, z_i^*), et les gains G_i.

Pour un signal aléatoire, les interférences entre branches parallèles du synthétiseur disparaissent, et il est possible de contrôler précisément

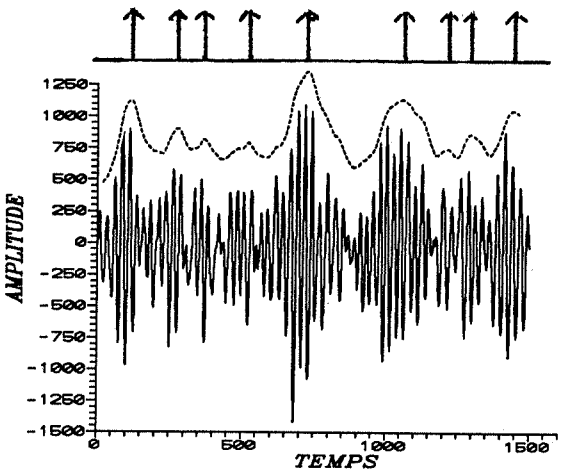


Figure 3: Bruit blanc Gaussien filtré (Lb = 100 Hz, Fc = 1000 Hz). On remarquera les maxima de son enveloppe temporelle et le processus ponctuel qui s'en déduit

l'enveloppe spectrale du signal de synthèse à l'aide des paramètres des résonateurs. En particulier, l'effet des zéros du spectre peut être simulé en jouant sur la largeur de bande des résonateurs. Les pôles et zéros de la fonction de transfert évoluent assez lentement. La description de la densité spectrale de puissance en termes formantiques, utilise pour chaque formant, l'amplitude, la largeur de bande et la fréquence centrale, qui sont des paramètres déterministes. Le type particulier de décomposition suivant les maxima formantiques, utilisée pour la synthèse, sera discutée dans la partie suivante. On considère, dans le paragraphe suivant, le signal d'une seule bande formantique.

2.3 Statistiques sur les maxima temporels de l'enveloppe du signal

Si l'on observe le signal issu de chaque bande formantique (soit un bruit blanc gaussien passé à travers un filtre passe-bande relativement étroit), les deux traits remarquables de ce signal sont la fluctuation aléatoire instantanée du signal et la fluctuation aléatoire de son enveloppe instantanée.

La figure 3 montre le signal issu du filtrage par un résonateur du second ordre, d'un bruit blanc Gaussien.

En utilisant la théorie de la modulation aléatoire, on peut représenter ce type de signal par les statistiques de son enveloppe temporelle instantanée et de sa fréquence instantanée. Nous allons citer des résultats dont les développements sont dans [9]:

1. étant donné un filtre passe bande, on peut calculer la densité de maxima λ_m de l'enveloppe E par unité de temps en fonction de la largeur de bande Bw du filtre:

$$\lambda_m = \alpha \times Bw \qquad (3)$$

Le coefficient α ne peut se calculer analytiquement que pour des formes de filtre particulières: il vaut 0.6411 pour un filtre passe-bande idéal (de gain rectangulaire), et 1.006 pour un filtre dont le gain est une gaussienne [8].

2. à partir des instants aléatoires d'apparition des maxima de l'enveloppe, on peut construire un processus ponctuel $\{t_i\}$.

3. la fréquence centrale du formant est la moyenne pondérée des fluctuations de la fréquence instantanée.

3 Synthèse du bruit par FOF aléatoires

La discussion précédente suggère d'utiliser un processus ponctuel aléatoire et une décomposition formantique pour synthétiser du bruit. Dans ce but, la méthode de synthèse proposée dans cet article utilise respectivement des impulsions de Poisson et des FOF.

3.1 Impulsions de Poisson

Un ensemble d'impulsions de Poisson $z(t)$ est défini par:

$$z(t) = \sum_i \delta(t - t_i) \qquad (4)$$

où t_i sont des points de Poisson (c-à-d un processus ponctuel obtenu par différentiation d'un processus de Poisson, voir [6]). $z(t)$ est un processus aléatoire stationnaire de moyenne $\eta_z = \lambda$, de fonction d'autocorrelation R_{zz} et de densité spectrale de puissance S_{zz}:

$$R_{zz}(\tau) = \lambda^2 + \lambda\delta(\tau) \ , \ S_{zz}(\omega) = 2\pi\lambda^2\delta(\omega) + \lambda \qquad (5)$$

Un ensemble de points de Poisson possède donc la même densité spectrale de puissance qu'un bruit blanc, excepté en $\omega = 0$, et réciproquement, la même fonction d'autocorrélation uniquement en $\tau = 0$. Le filtrage de points de Poisson par un filtre passe-bande parfait élimine la composante spectrale en $\omega = 0$. Alors, les statistiques du second ordre du processus $s(t)$ ainsi obtenu sont identiques à celles d'un bruit blanc filtré passe-bande. Le processus $s(t)$ ainsi obtenu à été introduit sous le nom de bruit de grenaille, et se rencontre dans de nombreuses situations physiques.

3.2 Bruit de grenaille

Soit un filtre de réponse impulsionnelle $h(t)$, le bruit de grenaille est défini comme le filtrage d'impulsions de Poisson, ou comme la somme des réponses impulsionnelles du filtre générées suivant un processus ponctuel de Poisson:

$$s(t) = \sum_i h(t - t_i) \qquad (6)$$

En général, le bruit de grenaille est défini en prenant pour $h(t)$ un filtre passe-bas. Pour notre propos, il faut supposer un filtre passe-bande qui élimine la composante $\omega = 0$. La densité spectrale de puissance et la moyenne du processus résultant valent:

$$S_{ss}(\omega) = \lambda |H(\omega)|^2 \quad , \quad \eta_s = \lambda \int_{-\infty}^{+\infty} h(t)dt \qquad (7)$$

de plus, si la densité λ est grande devant la durée de la réponse impulsionnelle $h(t)$, le processus $s(t)$ (bruit de grenaille de haute densité) est approximativement normal: ses statistiques du second ordre, mais également sa densité de probabilité sont alors similaires à celles d'un bruit blanc gaussien filtré. En résumé, les contraintes sur le filtre, pour synthétiser par bruit de grenaille un bruit blanc gaussien filtré, sont:

1. le filtre $h(t)$ doit être passe-bande: cela impose une densité spectrale de puissance identique à celle d'un bruit blanc filtré.
2. la largeur de bande du filtre, qui règle la durée et la décroissance de sa réponse impulsionnelle, doit être choisie assez petite devant la densité des points de Poisson pour que le signal filtré devienne gaussien.

Le paragraphe suivant discute du choix de $h(t)$.

3.3 Formes d'Ondes Formantiques

La décomposition spectrale en parallèle de l'équation 2 doit être considérée dans le domaine temporel, puisque ce sont les réponses impulsionnelles des filtres qui interviennent dans le bruit de grenaille (équation 4). Nous allons maintenant considérer des signaux discrêts, de période d'échantillonnage T. La réponse impulsionnelle $f_j(n, m)$ de la $j^{ième}$ section parallèle excitée à l'instant m vaut:

$$f_j(n, m) = A_j e^{-\alpha_j (n-m)T} \sin(\omega_j (n-m)T + \phi_j) \qquad (8)$$

où α_j règle la largeur de bande à -6 db du spectre d'amplitude, A_j l'amplitude temporelle (à une constante près le gain G_i de l'équation 2), ω_j la pulsation centrale, et ϕ_j la phase initiale du $j^{ième}$ formant. Un autre type de FOF permet de contrôler plus finement la forme de l'enveloppe spectrale indépendamment de la largeur de bande [10]. Pour notre propos, la phase ϕ_j peut être mise à zéro.

Le bruit peut être calculé comme un bruit de grenaille en définissant une excitation indépendante, avec un train virtuel d'impulsions de Poisson aux points $m_{ij}T$ différent pour chaque branche parallèle j:

$$x(n) = \sum_{j=1}^{M} \sum_i f_j(n, m_{ij}) \qquad (9)$$

Les durées des réponses impulsionnelles f_j sont réglées par les largeurs de bande des formants (suivant les paramètres α_j).

Le processus $x(n)$ devient gaussien si les densités λ_j associées aux points de Poisson dans chaque bande sont grandes comparées aux largeurs de bandes. En d'autres termes, chaque échantillon de signal doit être la somme d'un grand nombre de réponses impulsionnelles: comme les points de Poisson sont indépendants, leur sommation à travers les réponses impulsionnelles possède une densité de probablitité gaussienne. En résumé:

1. l'indépendance des impulsions de Poisson donne la blancheur de l'excitation virtuelle.
2. la sommation des impulsions (grâce a la durée des FOF) donne la densité gaussienne.
3. le spectre de puissance des FOF donne la densité spectrale de puissance.

Il est notable que seule la position de l'excitation virtuelle doit être aléatoire, ce qui est profondément différent du filtrage d'un bruit blanc, ou la position de l'excitation réélle est fixe (tous les échantillons), mais où l'amplitude de l'excitation est aléatoire.

L'analyse statistique des maxima de l'enveloppe d'un bruit gaussien montre que la largeur de bande de ce signal fixe, à une constante près, leur densité. La méthode de synthèse proposée nécessite, de même, de fixer la densité des points de Poisson, en accord avec la largeur de bande, afin de reconstituer un signal gaussien. Ces deux résultats sont en accord, mais il faut noter que la densité de maxima du bruit synthétisé est un ordre de grandeur plus faible que la densité de points de Poisson de synthèse: en effet, on se situe dans le cadre du bruit de grenaille de haute densité, pour lequel plusieurs impulsions d'excitation vont se combiner pour donner naissance à un maximum local de l'enveloppe du signal.

La partie suivante donne des résultats expérimentaux.

4 Résultats expérimentaux

On donne dans cette partie les résultats expérimentaux obtenus sur la synthèse de bruits continus: fricatives sourdes, voix chuchotée, bruit de souffle extrait d'une voix féminine.

4.1 Densité de probabilité

La densité λ des points de Poisson fixe le nombre de FOF qui s'additionnent pour chaque échantillon de signal, en fonction de la durée des réponses impulsionnelles, donc de la largeur de bande. La théorie

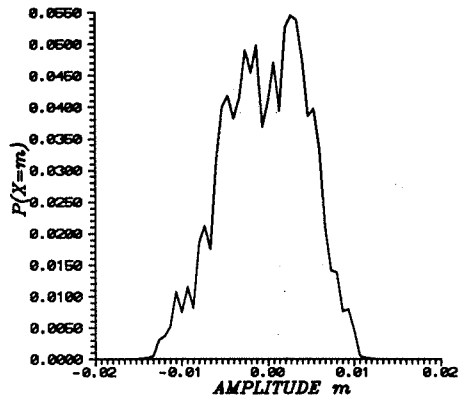


Figure 4: Densité spectrale de puissance pour un rapport λ/largeur de bande = 2.66

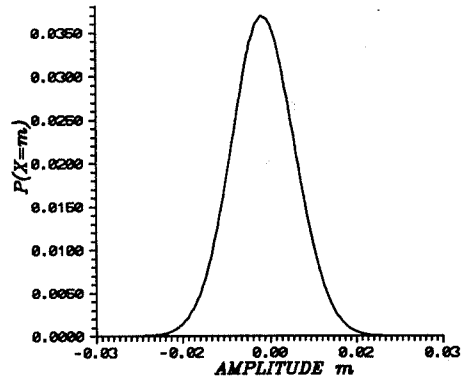


Figure 5: Densité spectrale de puissance pour un rapport λ/largeur de bande = 15

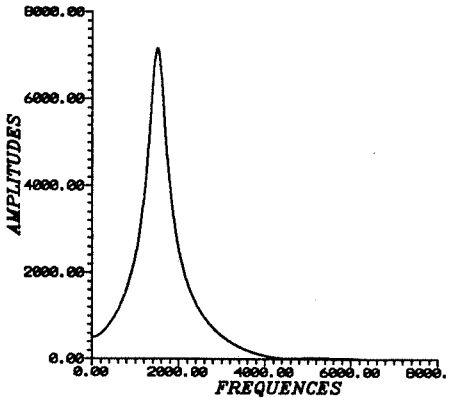


Figure 6: Spectre de puissance d'une FOF (Largeur de bande = 400 Hz, Fréquence = 1500 Hz)

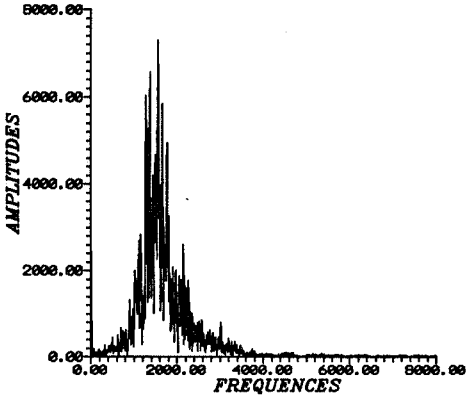


Figure 7: densité spectrale de puissance d'un signal de FOF aléatoires (Lb = 400 Hz, Fc = 1500 Hz)

prévoit que le bruit obtenu est gaussien lorsque la densité des points de Poisson est grande devant la largeur de bande des FOF. En pratique on peut considérer que le signal devient gaussien lorsque le rapport densité/largeur de bande devient supérieur à 10. Si les largeurs de bande sont exprimées en Hertz et les densités en points par seconde, le rapport densité/largeur de bande donne le nombre de FOF d'amplitude significative additionnées par échantillon. Les figures 4 & 5 illustrent les densités de probabilité empiriques obtenues pour des signaux à un seul formant, en faisant varier le rapport densité de points de Poisson/largeur de bande, en dessous et au dessus de la valeur critique, pour obtenir un bruit gaussien.

Enfin, les caractéristiques spectrales du bruit sont fixées par les paramètres formantiques (fréquence centrale, amplitude et largeur de bande des formants). Le signal de sortie obtenu est ainsi équivalent à du bruit blanc gaussien filtré.

4.2 Densités spectrales de puissance

Les densités spectrales de puissance des bruits de synthèse ont été calculées par périodogramme modifié.

Pour un signal à un formant, les figures 6 & 7 donnent respectivement le spectre d'amplitude d'une FOF et la densité spectrale de puissance du bruit correspondant.

4.3 Synthèse de fricatives

La synthèse est réalisée en deux étapes:

1. Un ensemble de points de Poisson est calculé à partir des densités définies pour chaque section paralèlle (la génération du processus poissonien est réalisée à partir d'un algorithme décrit dans [7]).

2. Les FOF sont générées a ces instants aléatoires à partir des paramètres formantiques déterministes (fréquences centrales, amplitudes et largeur de bande).

En prenant des densités relativement élévées, les signaux sont gaussien dans chaque bande formantique. Les figures 8 & 9 montrent les spectrogrammes d'un segment de parole naturelle [a$\int$a] et d'un segment de parole où la fricative est synthétisée suivant le principe précédemment décrit.

Une comparaison entre cette méthode et le synthétiseur de Klatt, à partir de la même analyse acoustique est reportée dans [2]. Il est apparu que

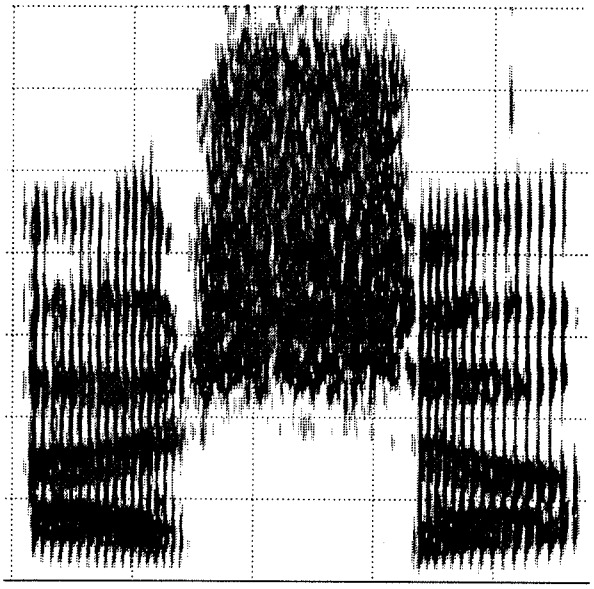


Figure 8: Spectrogramme du segment [a∫a] naturel

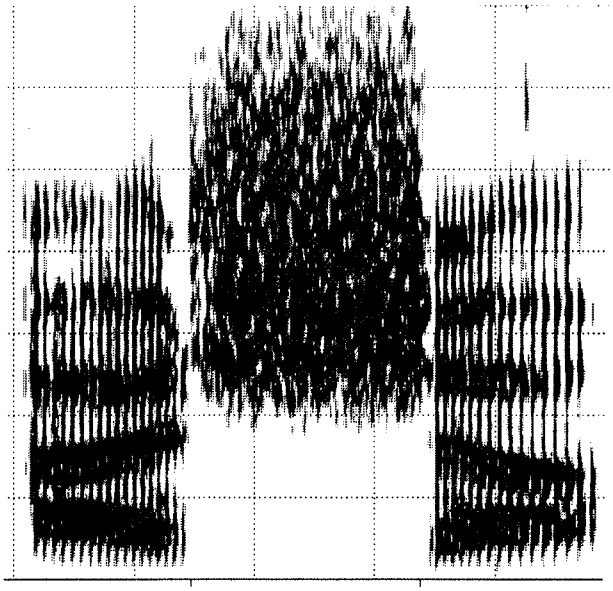


Figure 9: Spectrogramme du segment [a∫a] où la fricative est synthétique

l'intelligibilité est comparable, mais la synthèse par FOF aléatoires a été préférée quand au naturel.

4.4 Synthèse de souffle

Le bruit de souffle à été obtenu par un système d'analyse-synthèse qui sépare la partie quasi-harmonique de la partie aléatoire du signal de parole [3]. La synthèse à été effectuée automatiquement en utilisant l'analyse formantique décrite dans [1], et un ajustement des densités de points de Poisson en fonction des largeurs de bande.

5 Conclusion

La méthode présentée dans cet article n'est appliquée qu'aux bruits continus dont les propriétés statistiques évolent lentement dans le temps. L'extension que nous voulons apporter est l'application de cette méthode aux autres classes de bruits (bruits transitoires, consonnes vibrantes ...) pour lesquels la structure temporelle parait très importante. Notre méthode, qui utilise une description temporelle du signal, est alors particulièrement adaptée.

References

[1] d'ALESSANDRO, C. (1990). "Time frequency speech transformation based on an elementary waveform representation", **Speech Comm.**, Vol. 9, No 5/6, pp. 419-431.

[2] GRAU, S. (1991). "Comparaison entre le synthétiseur Klatt et un synthétiseur à formes d'ondes, **Notes et Documents LIMSI 91-13**.

[3] GRAU, S., d'ALESSANDRO C. (1992)., "Analyse-synthèse par décomposition de la partie déterministe et de la partie résiduelle du signal de parole", $19^{ième}$ **Journées d'Etudes sur la Parole.**

[4] KLATT, D. H. (1980). "Software for a cascade/parallel formant synthesizer" JASA, Vol. 67, No. 3.

[5] PAEZ, M. D., GLISSON, T. H. (1972) "MinimumMean Squared-Error Quantization in Speech,"**IEEE Trans. Comm.**, Vol. Com-20, pp. 225-230, Avril 1972.

[6] PAPOULIS, A. (1965). **Probability, Random Variables, and Stochastic Processes**, McGraw-Hill, pp. 239-240, 266-267, 284-285, 379-385.

[7] PRESS, W. H., FLANERY, B. P., TEUKOLSKY, S. A., VETTERLING, W.T. (1986) **Numerical recipes: the art of scientific computing**, Cambridge University PressGraw-Hill, pp. 239-240, 266-267, 284-285, 379-385.

[8] RICE, O. S. (1944-45). "Mathematical analysis of random noise", **Bell System Journal**, Vol. 24, pp. 282-332; Vol. 25, pp. 46-156.

[9] RICHARD, G., d'ALESSANDRO C., GRAU S. (1992). "Unvoiced speech analysis and synthesis using Poissonian random formant-wave-functions ", 6^{th} **European Signal Processing Conference.**

[10] RODET, X. (1980). "Time Domain Formant-Wave-Function Synthesis", in: J.C. Simon ed., **Spoken Language Generation and Understanding**, D.Reidel publishing compagny, Dordrecht.

[11] SHADLE, C. H. (1985), "The Acoustics of fricatives consonants", **Ph.D. thesis, MIT**, Cambridge, Mass. R.L.E. Technical Report 506.

19es J.E.P. - Bruxelles, 19 au 22 mai 1992.

MODELISATION DES VARIATIONS MICROMELODIQUES CO-INTRINSEQUES DES CONSONNES OCCLUSIVES DU FRANÇAIS POUR LA SYNTHESE PAR REGLES

Serge SANTI

Laboratoire "Parole et Langage", URA CNRS 261
29, Av R. Schuman, 13621 Aix-en-Provence, FRANCE

Résumé

Nous avons développé et appliqué deux règles micromélodiques à un corpus de dix-huit logatomes synthétiques générés par règles sur un synthétiseur à formants. Les bases théoriques de notre démarche, ainsi que nos motivations en fonction des domaines de la synthèse vocale aussi bien que d'une description plus générale du rôle linguistique des composantes de l'intonation sont exposées. Les règles et leur mise en oeuvre au sein du système de synthèse sont décrites et discutées.

1- INTRODUCTION: définitions et hypothèses

Dans le but d'améliorer la qualité de stimuli synthétiques élaborés dans le cadre d'une étude sur les phénomènes de coarticulation entre les consonnes occlusives du français et les voyelles cardinales /a,i,u/, nous avons décidé de vérifier si l'ajout de règles rendant compte des phénomènes microprosodiques inhérents à notre corpus avait une influence sur la perception de nos stimuli. Des tests de perception sont en cours de réalisation et devront démontrer le bien fondé de cette hypothèse. Nous nous proposons de décrire la démarche ayant conduit à l'élaboration de nos règles ainsi que leur mise en oeuvre au sein du synthétiseur.

Par **microprosodie** ou faits microprosodiques, nous entendons les variations de fréquence fondamentale, de durée et d'intensité des unités segmentales, imputables à la nature acoustique spécifique de ces unités ainsi qu'aux effets inhérents aux phénomènes de coarticulation entres ces dernières. Nous désignerons par **micromélodie** l'ensemble des phénomènes microprosodiques relatifs à la seule variation de la fréquence fondamentale.

Les phénomènes microprosodiques peuvent être considérés comme des variations non soumises au contrôle linguistique et obéissant à des contraintes imposées par la nature physiologico-acoustique de la parole. La dichotomie entre variable conditionnée et variable intentionnelle ne s'applique pas de manière rigoureuse en synthèse vocale où c'est plutôt, le plus souvent, la taille de l'unité de programmation (niveau suprasegmental, segmental, sous-segmental...) qui "supplante" le niveau linguistique. Le comportement séquentiel de la machine n'est pas calqué sur le comportement linguistique du locuteur humain.

Nous distinguerons deux catégories de variations microprosodiques en reprenant la terminologie énoncée par Di Cristo et Hirst (1986). Les variations **intrinsèques** caractérisent les phénomènes imputables aux effets acoustiques, articulatoires et aérodynamiques de la production des unités de type phonémique. Les modifications de la fréquence fondamentale résultant des phénomènes de coarticulation seront regroupées sous le terme de variations **co-intrinsèques**. C'est à ce dernier type de variation que vont se référer nos règles.

Notre but n'est pas de présenter ici les différentes théories portant sur les causes des phénomènes microprosodiques, d'autant qu'aucun consensus ne semble avoir été réalisé à ce propos (explications de type physiologique, aérodynamique, acoustique) (voir, pour un compte rendu de ces théories, Di Cristo 1982, 1985). Si l'existence de ces phénomènes (et donc l'intérêt d'une étude spécifique) semble attestée, et ne peut être mise sur le compte de biais expérimentaux (Di Cristo, Hirst 1986), pour de nombreux auteurs (voir Hirst 1987), il ne semble pas nécessaire de prendre en compte ces variations dans le cadre d'études sur les fonctions linguistiques de l'intonation. Ces derniers considèrent plutôt ces micro-variations comme un "bruit" perturbant la production et la perception de l'information prosodique. L'auditeur devrait, en quelque sorte, reconstituer mentalement la forme canonique des schémas prosodiques, corrigeant la distorsion introduite par la microprosodie (Di Cristo et Chafcouloff 1981, Rossi 1981). En accord avec ce principe, le logiciel de modélisation automatique de la courbe de fréquence fondamentale à partir d'une entrée de parole naturelle, mis au point par Hirst et Espesser (1991), va placer les point-cibles de la courbe modélisée en fonction d'un filtrage préalable de la microprosodie.

Une telle perspective sous entend tout de même, à

priori, une perception effective de ces phénomènes par l'auditeur. Nous pensons que le système auditif filtre effectivement la composante microprosodique mais que ce résidu est susceptible d'être réutilisé en tant qu'information sur la nature phonémique du support segmental. En conséquence, nous pouvons légitimement nous poser la question de savoir si la prise en compte de tels phénomènes est pertinente dans le décodage de la synthèse. De toute façon, la prise en compte d'un indice naturel supplémentaire, pertinent ou non-pertinent linguistiquement et/ou au niveau du décodage acoustico-phonétique, devrait tout au moins se révéler positif dans le cadre de la mise au point d'une voix de synthèse plus naturelle. Dans le cadre de notre étude, la question va se poser en ces termes: la présence de règles rendant compte de la microprosodie va-t-elle modifier la perception de nos stimuli ? Dans l'affirmative, de quel type sera ou seront ce ou ces modifications? Les études portant sur le rôle perceptif, au niveau segmental, des variations microprosodiques sont peu nombreuses. Nous citerons les travaux de Chistovich (1969), Haggard et al. (1970), Fujimura (1971), Larreur et Boë (1973), Massaro et Cohen (1976), Kohler (1982). Ces travaux vont dans le sens de l'utilité de la prise en compte des faits microprosodiques en synthèse vocale. Le rôle principal joué par ces derniers semble se situer dans le décodage du trait de voisement. Nous voyons donc apparaître que ce n'est pas seulement un souci d'efficacité et de performance qui est ici à l'origine de notre démarche mais la définition du rôle et des fonctions de la composante microprosodique dans la perception de la parole. Le processus mis en oeuvre dans notre modélisation de règles micromélodiques ainsi que l'évaluation perceptive qui en découlera constituent tout autant une analyse *par* la synthèse qu'une analyse *pour* la synthèse. La dichotomie relative au paramètre de fréquence fondamentale, séparant les variables intentionnelles (liées a l'organisation prosodique d'une langue donnée) et les variables non-intentionnelles (tout au moins en partie), dont fait partie, à notre sens, la microprosodie (ainsi que d'autres phénomènes du type "ligne de déclinaison", pauses respiratoires, etc ...), peut s'appliquer dans le cadre de la production de la parole synthétique. Nous devons cependant garder à l'esprit que toute modification d'un quelconque paramètre est l'oeuvre exclusive de l'opérateur, celui-ci se devant de "jouer le jeu" chaque fois que cela s'avère possible et/ou nécessaire. En effet, s'il peut s'avérer commode et économique de ne pas rendre compte de toutes les manifestations acoustiques de la parole réelle, les critères de sélection seront, dans tous les cas, la pertinence perceptive, la non-concurence avec l'efficacité du système de synthèse et, éventuellement, la cohérence avec un modèle de production et/ou de perception.

2- METHODOLOGIE

2-1- Le synthétiseur

2-1-1- Présentation générale et configuration

Nous avons utilisé un synthétiseur paramétrique à formants du type proposé par D. Klatt (1980). Ce type de synthétiseur simule les caractéristiques acoustiques de la parole, considérée comme le produit du spectre de source et de la fonction de transfert du conduit vocal.

Les paramètres de commande du synthétiseur sont calqués sur ce modèle (tab.1).

Ce synthétiseur a été utilisé dans sa configuration parallèle. Ce dernier est implanté sur un mini ordinateur MASSCOMP 5400 et est doté d'un environnement interactif (mutifenêtrage, menus déroulants, visualisation graphique des paramètres de commande, etc...) mis au point au Laboratoire Parole et Langage à Aix-en-Provence par R. Espesser.

- PARAMETRES RELATIFS A LA SOURCE
  - F0 : fréquence fondamentale
  - AV : amplitude de voisement
  - AF: amplitude de bruit
  - AVS: amplitude de voisement quasi-sinusoïdal
  - G0: gain global
- PARAMETRES RELATIFS AUX RESONATEURS
  - F1 ------> F6 : hauteur des formants
  - A1 ------> A6 : amplitude des formants
  - B1 ------> B6 : largeur de bande des formants

tab.1 : principaux paramètres de commande du synthétiseur

2-1-2- Mise en oeuvre de la composante prosodique et microprosodique

Avant d'exposer nos règles, il nous semble important de préciser la façon dont s'organise la paramétrisation de la courbe de fréquence de fondamentale au sein du synthétiseur.

L'élaboration du paramètre F0 se doit de respecter les dimensions traditionnelles de la prosodie: hauteur, intensité et durée. Ces trois dimensions peuvent être manipulées avec une grande souplesse. La durée est directement dépendante de l'organisation générale des paramètres de commande, l'unique contrainte s'y appliquant est la taille de la trame de 10 ms. L'intensité est liée, de façon absolue et en toute logique, au caractère voisé du segment considéré. Par conséquent, les paramètres AV et AVS conditionnent directement cette dimension, au contraire des autre paramètres de source "non-voisés" AF et AH. De fait, la fréquence fondamentale est contrôlée en tout et pour tout par le paramètre F0 qui fixe la position de chaque trame dans le spectre (au Herz prés). Il est important de noter que le paramètre F0 évolue, selon notre modélisation, indépendamment des autres paramètres. Des valeurs lui seront attribuées quel que soit le caractère voisé ou non voisé du segment, c'est à dire y compris lorsque AV et AVS sont à zéro. Cette non congruence éventuelle n'a évidement aucune incidence perceptive sur le résultat synthétique. Par contre, cette possibilité offre un double avantage. Il est, en effet, commode de ne pas avoir à se préoccuper du paramètre F0 à chaque re-initialisation des paramètres de source. De plus, d'un point de vue théorique, il semble logique de considérer la courbe de fréquence fondamentale comme un phénomène issu

d'une programmation continue, indépendante des unités segmentales. La programmation de l'opérateur, en synthèse, est alors calquée sur la programmation linguistique du locuteur, dans le cas de la parole naturelle. Cette conception de courbe de fréquence fondamentale sous-jacente n'est pas nouvelle et a déjà fait l'objet d'une argumentation solide de la part d'auteurs spécialistes en prosodie (Hirst 1987). De ce fait, l'interruption du voisement pendant la tenue (variation microprosodique intrinsèque), lors de la synthèse des consonnes occlusives sourdes, n'est pas directement modélisée par le paramètre de commande F0 et n'est pas traité, dans cette étude comme un phénomène microprosodique.

2-2- Corpus

Nous avons réalisé les synthèses de segments bi-syllabiques de type VCV où V est une des voyelles cardinales /a, i, u/ et C une des consonnes occlusives du français /b, d, g, p, t, k/, soit au total les 18 logatomes suivants (tab.2). Chaque logatome à été réalisé dans les deux "versions": avec micromélodie (AM) et sans micromélodie (SM).

/apa/	/ipi/	/upu/
/ata/	/iti/	/utu/
/aka/	/iki/	/uku/
/aba/	/ibi/	/ubu/
/ada/	/idi/	/udu/
/aga/	/igi/	/ugu/

tab.2 : liste des stimuli synthétisés dans les deux versions AM et SM

Notre corpus a été réalisé grâce à une procédure d'analyse/synthèse systématique d'un corpus de parole naturelle et à une modélisation des paramètres acoustiques pertinents issue du produit de notre analyse et des caractéristiques spécifiques du synthétiseur (Santi, 1989, 1990, 1991) et tiennent compte d'un test d'identification préalablement réalisé (Santi, Cavé, 1990). Les résultats de ce dernier ayant entraîné un certain nombre d'améliorations sur les logatomes les moins bien identifiés.

2-3- Règles micromélodiques, description et justifications

2-3-1- Schéma général et situation de la composante microprosodique en synthèse par règles

La décomposition de la composante prosodique, au sein d'un système de synthèse par règles (par opposition à un système par concaténation d'éléments pré-stockés ou la composante microprosodique est inclue au sein des unités constituantes) doit se faire (au minimum) en deux étapes. Le niveau supra-segmental est géré, du point de vue de la prosodie, par une panoplie de règles issues de contraintes contextuelles de type syntaxique, sémantique, pragmatique, etc...(niveau 1) Ces règles latentes sont filtrées par d'autre règles modélisant un certain nombre de contraintes de type phonotactiques (règles d'assemblages des "unités" prosodiques) ou physiologiques (pauses respiratoires, ligne de déclinaison,etc)(niveau 2). A ce stade est associé un premier jeu de points cibles de type macromélodique (MAC). Cette modélisation peut être injectée en l'état dans le synthétiseur ou complétée par une autre série de points cibles de type micromélodique (MIC) (niveau 3) (fig.1). Dans un souci de clarté, nous avons choisi de conserver les points cibles MAC et d'y ajouter les points cibles MIC plutôt que de modifier les premiers en fonction des seconds et de les combiner en une seule catégorie, ce qui est cependant tout à fait concevable.

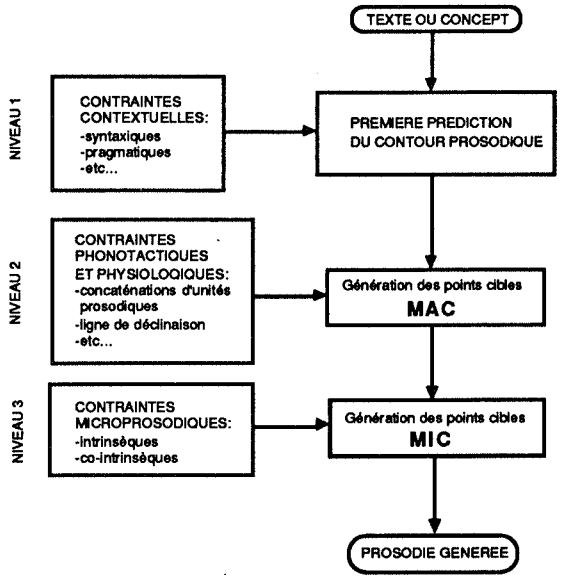


fig.1: schéma général de génération de la prosodie au sein d'un synthétiseur par règles

Nous considérons que la résultante des prédictions de la courbe de fréquence fondamentale effectuées aux niveaux 1 et 2 à donné, en sortie, un schéma mélodique unique pour l'ensemble de nos logatomes. S'agissant de mots isolés, ces prédictions sont réduites à leur plus simple expression, les contraintes régissant ces derniers étant très peu nombreuses. De ce fait, l'intonation de base de nos stimuli synthétiques reproduit globalement la courbe de fréquence fondamentale telle qu'elle a été produite lors de l'enregistrement des logatomes de parole naturelle ayant servi de support d'analyse à la modélisation acoustique sur laquelle se basent nos logatomes synthétisés. Nous n'avons volontairement pas localisé le point de changement d'orientation de la courbe de f0 sur l'une ou l'autre des syllabes constitutives de nos stimuli afin de ne pas conditionner trop précisemment la perception d'une syllabe accentuée (Kohler 1991). Cependant, la présence d'une durée systématiquement plus longue sur la voyelle V2 va dans le sens de la présence d'un relief accentuel (Duez 1987, Guaïtella 1988, Santi et Guaïtella 1990).

2-3-2- Règles micromélodiques

Deux règles micromélodiques ont été mises en oeuvre afin de rendre compte de certaines variations fréquentielles de type co-intrinsèque, produites lors de la coarticulation de nos voyelles et de nos consonnes. La première règle modélise les variations micromélodiques lorsque C est une consonne occlusive voisée, l'autre règle lorsque C est non-voisée. Aucune règle ou variation spécifique n'a été prise en compte à l'intérieur de chacune de ces sous-classes, aucune des différences relevées par les auteurs ayant essayé de quantifier et de systématiser les variations potentielles de ce type (Boë 1973, Di Cristo 1985) ne nous paraissant suffisante. Cette attitude, ainsi que le choix des valeurs temporelles et fréquentielles utilisées, ne sera pas argumentée ici mais fera l'objet d'une justification acoustico-perceptive lors de la discussion.

Avant de décrire ces règles dans le détail, il nous semble utile de préciser comment s'organisent et se calculent les coordonnées d'un point cible quelconque, Px. Px est défini par une coordonnée temporelle t (Px) et une coordonnée paramétrique f (Px) et se note Px [t (Px), f (Px)]. La valeur paramétrique f (Px) d'un point cible Px, inclus entre deux points cibles P1 et P2, ordonnés dans le temps et dont les coordonnées temporelles et paramétriques sont connues, s'obtient à partir des équations suivantes:

$$f(Px) = f(P1) + f(e)\,m$$

$$f(e) = \frac{f(P1) - f(P2)}{n}$$

$$n = \frac{|\,t(P1) - t(P2)\,|}{10} \qquad m = \frac{|\,t(P1) - t(Px)\,|}{10}$$

f (e) ---> valeur de variation paramétrique minimale entre deux trames
n ---> nombre de trames entre P1 et P2
m ---> nombre de trames entre P1 et Px

REGLE 1: CONSONNES VOISEES

Cette règle a pour conséquence d'abaisser la fréquence fondamentale lors de la tenue des consonnes occlusives voisées. La partie de la courbe comprise entre **déb.V1** et **fin.V1** reste inchangée. Un abaissement , de 20 Hz, se produit rapidement, en 10 ms (entre les points cibles **déb.MIC.VC** et **fin.MIC.VC**), la courbe évolue ensuite parallèlement à ce qu'était la courbe de base (en décalage de 20 Hz) jusqu'au point cible **déb.MIC.CV** ou elle récupère 20 Hz en atteignant le point **fin.MIC.CV**, toujours en 10 ms. Le schéma est alors identique à la courbe de base jusqu'au point **fin.V2** (fig.2). Aux trois points cibles originaux viennent s'ajouter quatre autres points cibles théoriques, en fait les point **fin.V1** et **déb.MIC.VC** sont confondus, ce dernier est simplement mentionné dans un souci de clarté et d'homogénéité dans l'exposition de la règle.

1e point cible: déb.V1 (0,120)
2e point cible: fin.V1 (100, 140) --------> déb.MIC.VC (100, 140)
3e point cible: fin.MIC.VC [t (déb.MIC.VC + Kt, f (déb.MIC.VC) - Kf] --------> fin.MIC.VC (110, 120)
4e point cible: déb.MIC.CV [t (fin.C) - Kt, f (fin.MIC.CV) - Kf]
5e point cible: fin.MIC.CV [t (fin.C), f (fin.MIC.CV)]
6e point cible: fin.V2 [t (fin.V2), 110]

Kt = 10 ms , Kf = 20 Hz

REGLE 2 : CONSONNES NON-VOISEES

Cette règle modifie la fréquence fondamentale de base en rehaussant de 20 Hz la F0 calculée par l'interpolation (**déb.MIC**), lors de la reprise du voisement. La fréquence chute alors brutalement, en 20 ms (2 trames), jusqu'à la valeur fréquentielle du point cible **fin.MIC** (fig .3).

Nous avons donc maintenant cinq points cibles au lieu de trois. L'ensemble des points cibles se reécrit de la façon suivante:

1e point cible: déb.V1 (0, 120)
2e point cible: fin.V1 (100, 140)
3e point cible: déb.MIC [t (fin.C), f (fin.C) + Kf]
4e point cible: fin.MIC [t (fin.C) + K't, f (fin.MIC)]
5e point cible: fin.V2 [t (fin.V2), 110]

K't = 20 ms , Kf = 20 Hz

La discontinuité de la courbe au point déb.MIC est un effet purement graphique, en effet une modélisation en points cibles reliés entre-eux par une interpolation ne peut être représenté de cette façon, cependant, cette représentation a été choisie en raison de son analogie avec une courbe "réelle" de la fréquence fondamentale.

3- DISCUSSION ET CONCLUSION

Malgré les nuances relevées par certains auteurs dans les schémas micromélodiques en fonction du lieu d'articulation des consonnes occlusives considérées (Larreur et Boë 1973), d'autres données, plus nombreuses (House et Fairbanks 1953, Lehiste et Peterson 1961, Mohr 1968, Di Cristo 1976), semblent ne contenir aucune variation significative à ce niveau. Nous pensons que, dans un premier temps tout au moins, notre attitude doit se circonscrire à l'étude de phénomènes attestés par l'analyse. N'oublions pas, non plus, le critère de "la plus grande simplicité possible", critère prédominant en synthèse. C'est là un des prix à payer si l'on veut envisager une intégration future de nos règles dans un système de synthèse à partir du texte.

Que ce soit pour la règle "voisé" ou la règle "non-voisé", les configurations de la fréquence fondamentale retenues correspondent tout à fait aux schémas relevés, sur des mots isolés ou sur de la parole continue, par nombre d'auteurs (voir Introduction). Cependant, si on compare les valeurs moyennes des variations de F0

imputables aux effets co-intrisèques des consonnes occlusives voisées et non-voisées relevées lors d'analyses acoustiques effectuées par nos prédécesseurs (voir Di Cristo 1982) et les valeurs fréquentielles que nous avons retenues, il apparaît que nous avons utilisé des valeurs sensiblement supérieures à ce à quoi l'on pourrait s'attendre. La plupart des variations micromélodiques issues de l'analyse tournent autour du seuil différentiel de perception des variations de fréquence fondamentale dans la parole (Rossi 1971, Rossi et Chafcouloff 1972). L'interprétation perceptive des faits micromélodiques peut alors se scinder en deux explications de type diamétralement opposé: ces variations, issues de contraintes incontournables, sont maintenues en deça du seuil afin, par exemple, de ne pas interférer avec le niveau supra-segmental, ou bien ces variations sont maintenues au delà du seuil afin de permettre une intégration perceptive (quelle qu'elle soit) de l'information acoustique ainsi générée. Notre opinion est, bien entendu, en faveur de cette seconde hypothèse. La synthèse offre l'avantage de pouvoir contrôler la dimension microprosodique en dehors des contraintes de production de la parole naturelle, l'occasion nous était donc donnée de tester indirectement l'hypothèse de l'intégration perceptive. En effet, le "problème du seuil" étant écarté par le choix de valeurs supérieures à ce dernier (tout en restant cohérentes), les résultats de tests de perception adéquats seront interprétables au niveau strict du décodage acoustico-phonétique. Les résultats devraient se révéler intéressants, par voie de conséquence, non seulement pour la synthèse vocale, mais également dans l'optique de la prise en compte des paramètres prosodiques dans la reconnaissance automatique de la parole. Par contre, l'interférence éventuelle avec le niveau prosodique suprasegmental ne pourra être testée ainsi (sur des logatomes), mais seulement sur de la parole continue. Une série de tests de perception est en cours de réalisation (Santi 1992, Santi et Cavé -à paraître-) qui devront permettre d'apporter des réponses à ces questions. Nous espérons que notre démarche permettra de relancer l'approche d'une conception de la microprosodie sortie du cadre strict de l'étude des fonctions suprasegmentales de l'intonation et replacée au niveau d'un traitement global de bas niveau de l'information acoustique.

BIBLIOGRAPHIE:

BOE J. L., 1973, "Etude de l'interaction source laryngienne - conduit vocal dans la détermination des caractéristiques intrinsèques des consonnes du français", *Bulletin de l'Institut de Phonétique de Grenoble*, vol.2, 1-24.

CHISTOVICH L. A., 1969, "Variation of the fundamental voice pitch as a discriminatory cue for consonants", *Soviet Physics Acoustics*, vol.14(3), 372-378.

DI CRISTO A., 1976, "Des traits acoustiques aux indices perceptuels. Application d'un modèle d'analyse prosodique à l'étude du vocatif en français", *Travaux de l'Institut de Phonétique d'Aix-en-Provence*, vol. 3, 213-358.

DI CRISTO A., 1982, *Prolégomènes à l'étude de l'intonation. Micromélodie*, Editions du CNRS, Paris.

DI CRISTO A., 1985, *De la microprosodie à l'intonosyntaxe*, Thèse d'état, Aix en Provence.

DI CRISTO A., CHAFCOULOFF M., 1981, "L'intonème progrédient en français: caractéristiques intrinsèques et extrinsèques", *Studia phonetica*, vol.18, 39-51.

DI CRISTO A., HIRST D., 1986, "Modelling french micromelody: analysis and synthesis", *Phonetica*, 43, 11-30.

DUEZ D., 1987, *Contribution à l'étude de la structuration temporelle de la parole en français*, Thèse de Doctorat d'Etat, Unioversité d'Aix-en-Provence.

FUJIMURA O., 1971, "Remarks on stop consonant synthesis experiments and acoustic cues"; in Hammonds, *Form and substance*, Akademisk Forlag, Copenhagen, 221-232.

HAGGARD M., AMBLER S., CALLOW M., 1970, "Pitch as a voicing cue", *J. Acoust. Soc. Am*, vol.47, 613-617.

HIRST D. J., 1987, *La représentation linguistique des systèmes prosodiques: une approche cognitive*, Thèse d'Etat, Aix-en-Provence.

HIRST D. J., ESPESSER R., 1991, "MOMEL mode d'emploi", *Document interne du Laboratoire Parole et Langage*, Aix-en-Provence.

HOUSE A. S., FAIRBANKS G., 1953, "The influence of consonant environment upon the secondary acoustical characteristics of vowels", *J. Acoust. Soc. Am*, vol.25(1), 105-113.

GUAITELLA I., 1988, "Variation de durée en syllabe accentuée", *Travaux de l'Institut de Phonétique d'Aix-en-Provence*, vol.12, 185-204.

KLATT D, 1980, "Software for a cascade/parallel synthesizer", *J. Acoust. Soc. Am.*, 67(3), 971-995.

KOHLER K., 1982, " f0 in the production of fortis and lenis plosives", *Phonetica*, vol.39, 199-218.

KOHLER K., 1991, "A model of German intonation", *conférence*, Aix-en-Provence, le 03/04/1991.

LARREUR D., BOE J. L., 1973, "Etude de l'influence des variations de la fréquence laryngienne sur l'intelligibilité et la qualité des consonnes générées par vocodeur", *Bulletin de l'Institut de Phonétique de Grenoble*, vol.2, 103-126.

LEHISTE I., PETERSON G. E., 1961, "Some basic considerations in the analysis of intonation", *J. Acoust. Soc. Am.*, vol.33(4), 419-425.

MASSARO D. W., COHEN H. M., 1976, "The contribution of fundamental frequency and voice onset time to the /zi/ - /si/ distinction", *J. Acoust. Soc. Am.*, vol.60(3), 704-717.

MOHR B., 1968, "Intrinsic fundamental frequency variation: II ", *Monthly Internal Memorandum*, Phonol. Lab., University of California, June, 23-32, (cité par Di Cristo 1982).

ROSSI M., 1971, "Le seuil de glissando ou seuil de perception des variations tonales pour les sons de la parole", *Phonetica*, vol.23, 1-33.

ROSSI M., 1981, "Prosodical aspects of speech production"; in Ferrero, *Proceedings of 4th FASE Symposium on Acoustics and Speech*, vol.2, 125-157.

ROSSI M., CHAFCOULOFF M., 1972, "Recherches sur le seuil différentiel de fréquence fondamentale dans la parole, *Travaux de l'Institut de Phonétique d'Aix-en-Provence*, vol.1, 179-185.

SANTI S., 1989, "Extraction et modélisation de logatomes synthétiques pour la synthèse du français", *Travaux de l' Institut de Phonétique d'Aix-en-Provence*, vol.13.

SANTI S., 1990, "Modeling and identification of VCV synthetic nonsense-words in French", *Proceedings of LP'90 Conference*, Prague. (sous-presse)

SANTI S., 1991, "Effets de règles micromélodiques sur la perception de logatomes synthétiques", *Proceedings of the XIIth International Congress of Phonetic Sciences*, Aix-en-Provence.

SANTI S., 1992, *Synthèse de la parole: modélisation acoustique et évaluation perceptive*, Thèse de Doctorat (en préparation).

SANTI S., CAVE C., 1990, "Evaluation de l'intelligibilité de la synthèse: comparaison entre parole synthétique et parole naturelle", *Actes du Premier Congrès Français d'Acoustique*, Les Editions de Physique, 475-479.

SANTI S., CAVE C., (à paraître), "Perceptual data on the role played by micromelody in the identification and discrimination of synthetic speech segments".

SANTI S., GUAITELLA I., 1990, "Variations of duration in stressed syllables taken from French read sentences, *J. Acoust. Soc. Am.*, vol.87 (s1).

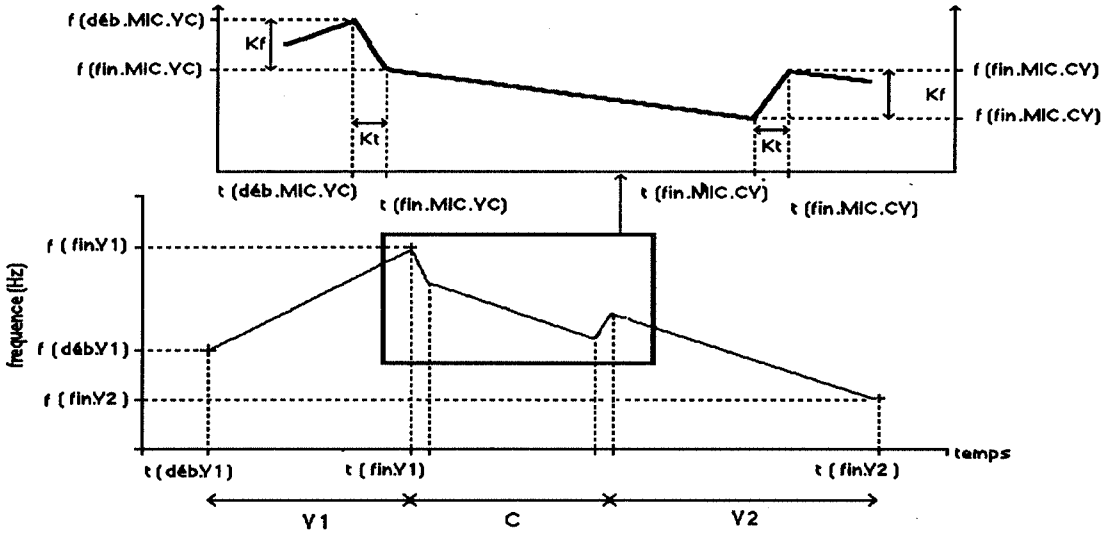


fig.2: contour mélodique de base modifié par la règle micromélodique "voisé".

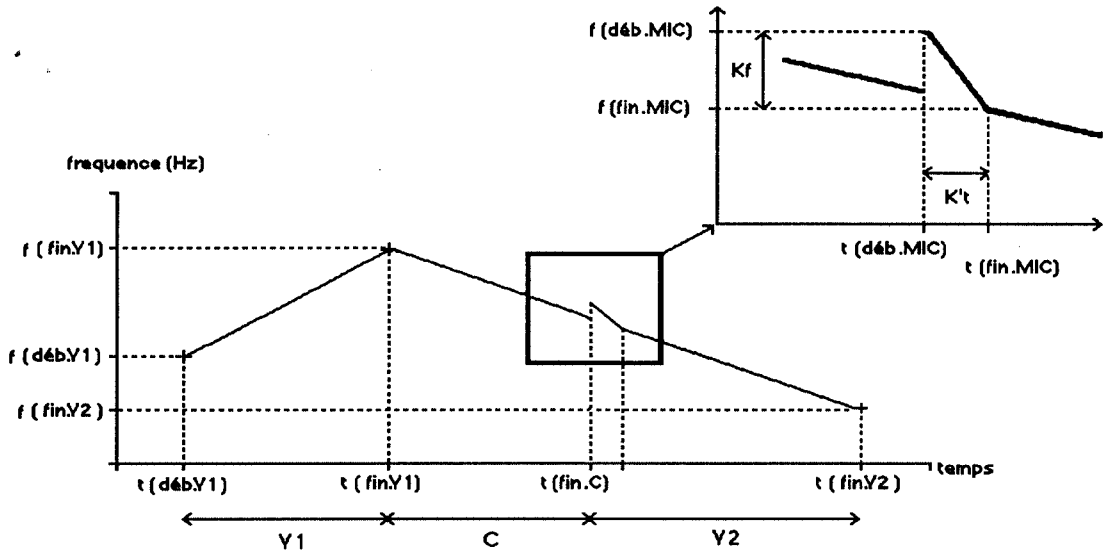


fig.3: contour mélodique de base modifié par la règle micromélodique "non-voisé".

19es J.E.P. - Bruxelles, 19 au 22 mai 1992.

SYNTHESE PAR REGLES DES GROUPES CONSONANTIQUES DU FRANÇAIS

Martine Garnier-Rizet

LIMSI-CNRS BP 133 91403 ORSAY CEDEX

Résumé

Le propos de cette étude est la synthèse, par formants, des groupes consonantiques du français (GC) à partir de règles acoustico-phonétiques. Un inventaire des GC du français est établi suivi d'un classement en plosives(P), fricatives(F), liquides(L), nasales(N). La description acoustique des 16 groupes de GC souligne, pour la synthèse, l'importance des phénomènes suivants : l'organisation temporelle du GC ; les phases acoustiques de ses constituants ; les mouvements fréquentiels, les discontinuités et les variations d'amplitude à la frontière C_1/C_2 ; l'incidence sur les fréquences et sur le type de source d'excitation, d'une consonne sur la totalité d'une phase au moins de l'autre consonne ; les effets de coarticulation du GC avec les voyelles adjacentes. Des exemples de synthèse et des règles déclenchées sont donnés.

1 INTRODUCTION

Le propos de cette étude est la synthèse, par formants, des groupes consonantiques (GC) du français à partir de règles acoustico-phonétiques. Les règles, opérant sur une chaîne phonémique, partent de valeurs cibles pour chaque phonème, stockées dans un tableau, et génèrent un fichier de paramètres lu par le synthétiseur, proche du modèle de Klatt dans sa version hybride. Lorsqu'aucune règle n'est appliquée, la synthèse d'un groupe consonantique est simplement la synthèse d'une suite de deux phonèmes à partir de leurs seules valeurs cibles de formants, largeurs de bande, amplitudes, durée etc... stockées. Ainsi, le contenu des règles de synthèse tient compte de : l'organisation temporelle du GC, la structure acoustique de ses constituants, la phase de transition interne C_1C_2, l'effet de coarticulation du GC avec l'environnement vocalique. L'effet de réduction de la durée des consonnes dans un GC et la prédiction de la durée du GC suivant sa position dans le mot s'inspire d'études déjà réalisées, (Bartkova et Sorin, 1987), (O'Shaughnessy, 1981), (Nishinuma, Duez, Paboudjian, 1991). Notre propos est donc de présenter dans une première partie l'inventaire des GC du francais, le corpus utilisé et son classement. La deuxième partie est consacrée à l'étude, par classe, de la structure acoustique du GC, de sa phase de transition interne et de son incidence sur les voyelles adjacentes, ainsi qu'aux règles de synthèse qui ont été développées.

2 INVENTAIRE DES GROUPES CONSONANTIQUES DU FRANÇAIS.

Un inventaire des groupes consonantiques du français a été établi en effectuant un comptage des suites CC d'un texte phonétisé automatiquement par règles, (Prouts, 1980). Ce texte est la transcription des séances des questions du Sénat en 1988-1989 et contient environ un million de mots. Trois restrictions ont été faites à partir de la liste obtenue :

- Seuls les groupes de deux consonnes sont pris en compte.
- Seuls les GC situés entre les frontières du mot, et non à la jointure de deux mots, sont retenus.
- Les semi-voyelles [w, u, j] ne figurent pas dans les GC étudiés.

Les GC sont regroupés avec le classement suivant : plosives (P), fricatives (F), liquides (L), nasales (N). Ainsi, 16 groupes sont définis : PP, PF, PL, PN, FP, FF, FL, FN, LP, LF, LL, LN, NP, NF, NL, NN.

▪ Les semi-voyelles [w, u, j] ne figurent pas dans les GC étudiés.
Les GC sont regroupés avec le classement suivant : plosives (P), fricatives (F), liquides (L), nasales (N). Ainsi, 16 groupes sont définis : PP, PF, PL, PN, FP, FF, FL, FN, LP, LF, LL, LN, NP, NF, NL, NN.
Un corpus de logatomes aC_1C_2a a été enregistré par un locuteur masculin ainsi qu'un corpus de mots disyllabiques ou trisyllabiques où le GC est situé en position intervocalique.
La liste à partir de laquelle a été constituée le corpus contient les GC les plus fréquents donnés par (Aubergé, Boë, Lefèvre, 1988) et au moins 1 représentant de chacun des 16 groupes. La distribution des GC, par groupe est donnée dans le tableau 1.

	P N	F	L	
P	6	15	12	11
F	7	4	12	11
L	12	12	2	4
N	3	1	4	2

TAB.1 DISTRIBUTION DES GC PAR GROUPE

3 ANALYSE ACOUSTIQUE DESCRIPTIVE

Par convention, dans une séquence $V_1C_1C_2V_2$, V_1 est la voyelle précédent le GC, V_2 celle qui le suit. Pour indiquer les effets de coarticulation des consonnes avec les voyelles V_1 et V_2 on note :

▪ V_1C_1 et C_2V_2, lorsque l'effet de la coarticulation s'exerce uniquement entre la consonne et la voyelle la plus proche.

▪ $V_1 - C_2$ et $C_1 - V_2$, lorsque l'effet de la coarticulation s'exerce également entre la consonne et la voyelle qui en est plus éloignée.

On a choisi de décrire dans l'ordre :

▪ les phases acoustiques qui composent le GC.

▪ les mouvements fréquentiels à la frontière C_1/C_2, les discontinuités et les variations d'amplitude. On appelle transition interne l'ensemble de ces phénomènes.

▪ l'incidence sur les fréquences, et sur le type de source d'excitation, d'une consonne sur la totalité d'une phase au moins de l'autre consonne.

▪ les effets de coarticulation des consonnes avec les voyelles adjacentes, tels qu'ils ont été définis précédemment.

a/ Groupe P_1P_2

Le GC P_1P_2 se caractérise par trois ou quatre phases acoustiques : la tenue suivie en général de l'explosion de P_1 puis la tenue suivie de l'explosion de P_2. L'absence de l'explosion de P_1 ne semble pas importante perceptivement, on a donc choisi de toujours modéliser les quatre phases acoustiques du GC.
On n'observe aucune transition interne, et pas d'incidence non plus de P_2 sur la répartition fréquentielle de l'explosion de P_1 et réciproquement.
La coarticulation de P_1 et P_2 s'exerce avec la voyelle la plus proche uniquement.

La synthèse des séquences $V_1P_1P_2V_2$ se réduit, si l'on ne considère pas les réductions de durée, à une concaténation de V_1P_1 et P_2V_2.

b/ Groupes P_1F_2 et F_1P_2

Les GC P_1F_2 et F_1P_2 sont composées de trois phases acoustiques, la tenue et l'explosion de P_1 suivie du bruit fricatif F_2 ou le bruit fricatif F_1 suivi de la tenue et de l'explosion de P_2.

Dans ce dernier cas, pour les F_1P_2 non voisées, on note une discontinuité en fréquence et en amplitude due à l'interruption brutale du bruit avant la tenue de la plosive. On observe un temps de transition pendant lequel la limite inférieure du bruit de F_1 tend vers la fréquence d'explosion de P_2, c'est le cas pour la séquence [afta] par exemple. Pour les F_1P_2 voisées, la discontinuité est fréquentielle et on observe une variation de l'amplitude de voisement de la fricative à la plosive.
Pour P_1F_2, l'explosion de P_1 se confond avec le début du bruit de F2, on a également une transition interne réalisée par la limite inférieure de bruit de F_2.
Le lieu d'articulation de la fricative a une incidence sur la répartition fréquentielle de l'explosion de la plosive, plus basse que sa valeur cible et plus compacte pour [t], par exemple, devant f, dans [atfa]. Ce phénomène

de coarticulation doit être modélisé avec exactitude dans les règles qui modifient les paramètres formantiques et les valeurs d'amplitude, son absence entraîne une mauvaise identification de la plosive. Il est préférable, en effet, de ne pas synthétiser d'explosion plutôt que de modéliser une mauvaise répartition des fréquences.

Pour les GC non voisés, il semble que la coarticulation se fasse avec la voyelle la plus proche, du type V_1C_1 et C_2V_2. En revanche, pour les séquences P_1F_2 et F_1P_2 voisées, on note une coarticulation P_1 - V_2 et V_1 - P_2 particulièrement marquée avec la fricative [v] et les plosives [d, g]. Comme on peut le constater sur la Fig.1 le locuteur réalise le [v] de façon très voisée et l'on peut observer le suivi des formants de la voyelle. Cet effet de coarticulation nécessite d'ajouter des règles qui s'appliquent sur les voyelles, en prenant en compte un contexte de deux consonnes.

c/ Groupes P_1L_2 et L_1P_2

Les séquences P_1[R] se caractérisent par la tenue puis l'explosion de P_1 suivie du bruit du [R] lorsque la plosive est non voisée, suivie du [R] voisé avec une structure formantique, lorsque la plosive est voisée. En revanche, [l] a toujours été réalisé voisé.
La transition interne est réalisée par les mouvements de F2 et F3 du [l] et du [R] voisé qui tendent vers la fréquence d'explosion de la plosive. La discontinuité en fréquence est nette dans le cas L_1P_2. Notre locuteur réalise parfois un bruit d'explosion situé juste avant la tenue de la plosive dont nous n'avons pas tenu compte dans nos règles, car il n'apparait pas systématiquement et peut être dû à une variante individuelle.
L'amplitude est croissante dans le cas P_1L_2 voisée et décroissante dans le cas L_1P_2 voisée. Lorsque la plosive n'est pas voisée, [R] est réalisé avec très peu d'énergie surtout en position C_1.
L'identification du [l] synthétique est liée au respect de la transition de la voyelle avec [l] pour les voyelles postérieures. Dans ce cas, les phénomènes de coarticulation sont plutôt du type V_1C_1 et C_2V_2. Pour les voyelles antérieures, la coarticulation s'exerce entre la plosive et la voyelle adjacente au [l]. Lorsque [R] est voisé, la coarticulation se fait entre la plosive et les deux voyelles V_1 et V_2, cette coarticulation bien visible dans les séquences [agra] et [arga].

Le nombre de règles ajoutées pour modéliser les GC P_1L_2 et L_1P_2 est élevé puisqu'il comprend : les modifications du type de source d'excitation, bruit ou voisement, pour le [R], les transitions fréquentielles de [l, R], et celles des voyelles adjacentes.

d/ Groupe P_1N_2 et N_1P_2

Le groupe P_1N_2 est constitué de quatre phases acoustiques, la tenue et l'explosion de la plosive, un temps de tenue voisé ou non suivi du murmure nasal de la nasale. Nous avons tenu compte du temps de tenue de la nasale dans nos règles car on le retrouve systématiquement chez notre locuteur dans les GC P_1N_2. Ceci entraîne, dans le module de règles, des modifications au niveau de la segmentation de la chaîne phonémique en unités acoustiques plus petites.
On ne distingue aucune phase de transition interne en fréquence. Lorsque la plosive est voisée, la barre de voisement est continue de la plosive à la nasale.
Le lieu d'articulation de la nasale a une incidence sur la répartition fréquentielle de l'explosion de la plosive ; l'explosion du [t] est plus basse devant [m] par exemple.
La discontinuité en fréquence étant marquée entre les nasales et les voyelles, les effets de coarticulation de P_1N_2 et N_1P_2 s'exercent essentiellement entre la plosive et la voyelle la plus proche.

La nasale [m] est difficile à synthétiser et d'autant plus en position C_2 de GC, où elle est encore trop souvent identifiée comme [b].

e/ Groupe F_1F_2

Les séquences F_1F_2 sont constituées de deux phases acoustiques, le bruit de F_1 suivi du bruit de F_2 auxquels s'ajoute le voisement lorsque les fricatives sont voisées. On constate une variation de l'amplitude de bruit ou de voisement avec un maximum d'amplitude marqué à la frontière des deux consonnes. On n'observe pas de transition interne significative, en revanche, une incidence de F_1 sur la limite inférieure du bruit de F_2 : la limite inférieure de [s] est plus basse devant [f] par exemple dans [asfa]. Dans le cas de fricatives voisées de lieu d'articulation éloigné, [z, v], on a une

continuité des trois premiers formants de V_1 jusqu'à V_2, le bruit [z] apparaissant au-delà, dans les hautes fréquences.
Il n'y a pas de phénomène de coarticulation significatif entre les voyelles et le GC non voisé F_1F_2. La coarticulation est plutôt du type V_1C_1 et C_2V_2.

Les règles portent sur les modifications de fréquences centrales et de largeurs de bande des fricatives, sourdes et voisées, les durées de transition et les valeurs de formants des voyelles.

f/ Groupe F_1L_2 et L_1F_2

On observe une transition interne qui se caractérise par un mouvement de F2 et F3 du [l] vers la fricative.
[R] a très peu d'énergie lorsqu'il est bruité, c'est-à- dire avant ou après une fricative sourde et c'est suivi ou précédé de [v] qu'il est le plus voisé.
Lorsque [R] est voisé, on note une continuité des trois premiers formants de V_1 jusqu'à V_2, le bruit de la fricative apparaissant dans les hautes fréquences, dans la séquence [arza] par exemple.

Les modifications des règles portent sur le rapport entre l'amplitude de bruit et l'amplitude de voisement pour les fricatives, la durée de transition du [l], les durées de transition et les valeurs de formants des voyelles.

g/ Groupe F_1N_2 et N_1F_2

Les phases acoustiques du groupe F_1N_2 sont le bruit de la fricative suivi d'un temps de tenue, voisé ou non et du murmure de la nasale. On retrouve dans ce groupe les deux phases tenue et murmure nasal rencontrées dans le groupe P_1N_2, ce qui est peut-être une caractéristique de notre locuteur. Lorsque la fricative est voisée, il n'y a pas de discontinuité dans la barre de voisement mais une variation croissante d'amplitude de la fricative vers la nasale.
Il apparait une phase de transition interne pendant laquelle la limite inférieure du bruit de la fricative descend vers la nasale, dans la séquence [asma] par exemple.

h/ Groupe L_1L_2

Ce groupe est composé de deux phases acoustiques, où [l] est toujours voisé et [R] parfois réalisé voisé, parfois non voisé et ce dans un contexte identique. Lorsque [R] est voisé, on observe une discontinuité des formants à la frontière des deux consonnes.

La synthèse de ces deux GC [lR] et [Rl] a été réalisée avec un [R] voisé et un [R] bruité. Le GC avec [R] voisé est perçu plus naturel.

i/ Groupe L_1N_2 et N_1L_2

Deux phases acoustiques constituent le groupe L_1N_2 dans lequel [R] est réalisé parfois voisé, parfois non voisé. On constate une discontinuité des fréquences et pas de mouvement de transition.
Il semble qu'il y ait un maximum d'amplitude du signal à la frontière des deux consonnes.

Les règles portent sur l'évolution de l'amplitude de bruit ou de voisement de [R] et sur ses valeurs de fréquences centrales.

j/ Groupe N_1N_2

On observe une continuité des formants en basse fréquence, et une discontinuité en hautes fréquences. On observe peu de mouvements de coarticulation avec les voyelles adjacentes.

Les figures 2, 3 et 4 présentent les spectrogrammes des synthèses de [afta], [akma] et [agra] avec, pour ce dernier, un exemple de quelques règles déclenchées.

4 CONCLUSION

La synthèse par règles des GC du français repose sur la modélisation des phénomènes fond-amentaux suivants :

- la réduction et parfois l'allongement de la durée des consonnes C_1C_2 du GC.

- les évolutions en fréquences et en amplitude des liquides suivies ou précédées de plosives ou de fricatives.

- l'incidence des fricatives sur la répartition fréquentielle de l'explosion des plosives.

- l'introduction d'une phase supplémentaire de tenue pour les nasales précédées par une plosive ou une fricative.

- les phénomènes de forte coarticulation de type $V_1 - C_2$ et $C_1 - V_2$ lorsque C1 et C2 sont des plosives et [R] ou [l] la deuxième consonne du groupe.

Si l'on devait effectuer un classement suivant le nombre de règles appliquées pour la synthèse de chaque groupe, les groupes PL, PF, FL arrivent en tête, ce sont les plus fréquents en français. Les groupes PP et NN arrivent en fin de liste, ils sont d'une part moins nombreux, et d'autre part moins fréquents.

Un premier ensemble de règles a été développé pour la synthèse des groupes consonantiques. Il a permis de constituer un corpus de test de logatomes aC_1C_2a dont les résultats, en cours d'évaluation, vont permettre de comparer le score d'identification de chacune des consonnes en groupe consonantique et en contexte vocalique.

REFERENCES

Aubergé, V., Boë., L-J., Lefèvre, J-P. (1988), *Lexiques et Groupes Consonantiques*, Actes des 17èmes Journées d'Etudes sur la Parole, Nancy, 20-22 septembre, pp.55-60

Bartkova, K., Sorin, Ch., (1987), *A Model of Segmental Duration for Speech Synthesis in French*, Speech Communication, Vol. 6, N°3 pp. 245-260

Meunier, Ch. (1990), *Groupes Consonantiques : Premier Inventaire des Réalisations Acoustiques des Phases de Transition,* Actes des 18èmes Journées d'Etudes sur la Parole, Montréal, 28-31 mai, pp.69-72

O'Shaughnessy, D., (1974) *Consonant Durations in Clusters*, IEEE Trans. Acoust. Speech Signal Process., Vol. ASSP 22, pp.282-295.

Prouts, B., (1980), *Contribution à la synthèse de la parole à partir du texte ; transcription graphème-phonème en temps réel sur microprocesseur,* Thèse de Docteur-ingénieur, Université Paris-Sud, novembre 1980.

Rochette, C. (1974), *Les Groupes de Consonnes en Français : étude de l'enchaînement articulatoire à l'aide de la radiocinématographie et de l'oscillographie* (Klincksieck, Paris)

Rossi, M. (1968), *Au sujet des Groupes Consonantiques du Français,* Revue d'Acoustique, Vol. 3 pp.306-311

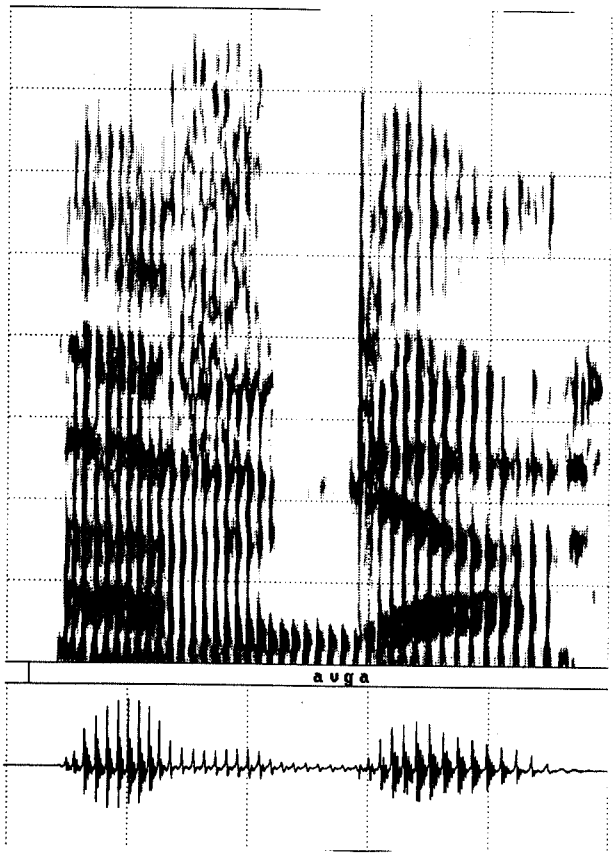


Fig. 1 [avga] naturel

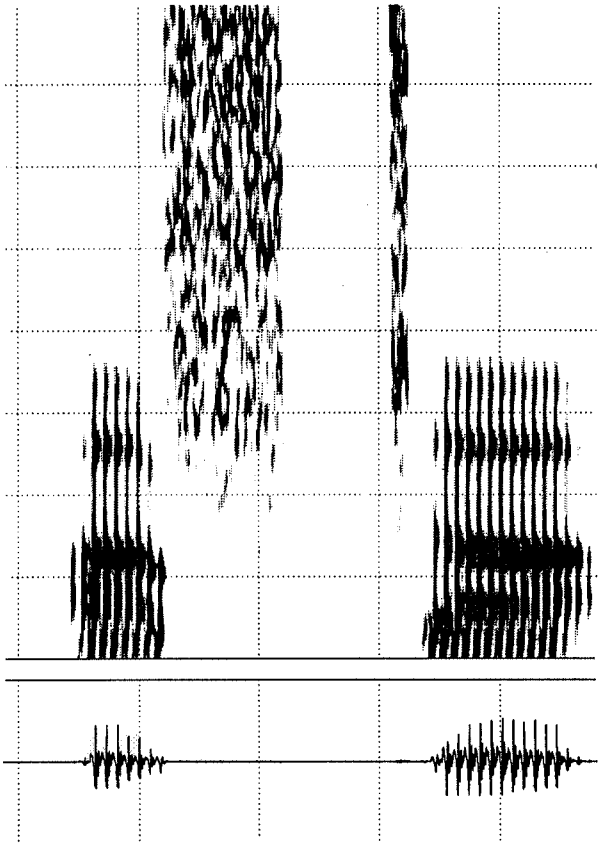


Fig. 2 [afta] synthétique

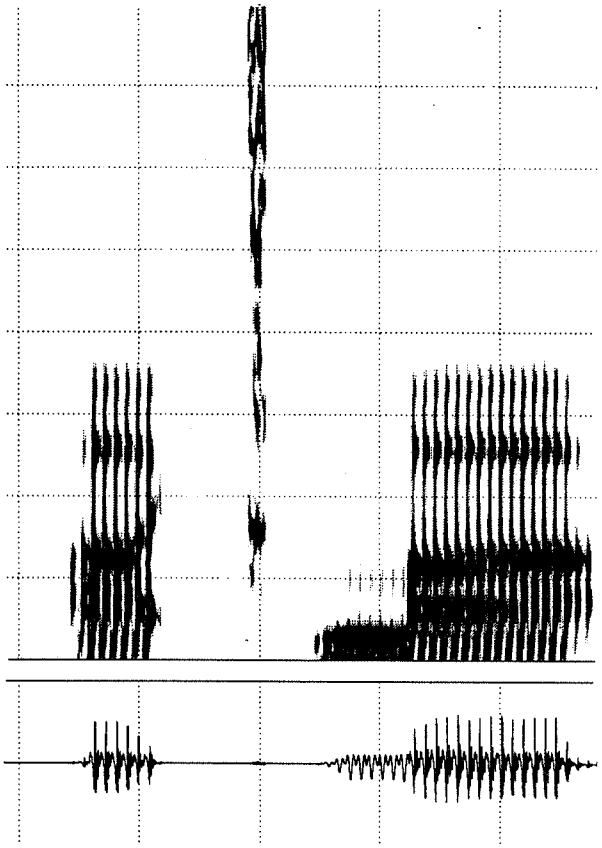


Fig. 3 [akma] synthétique

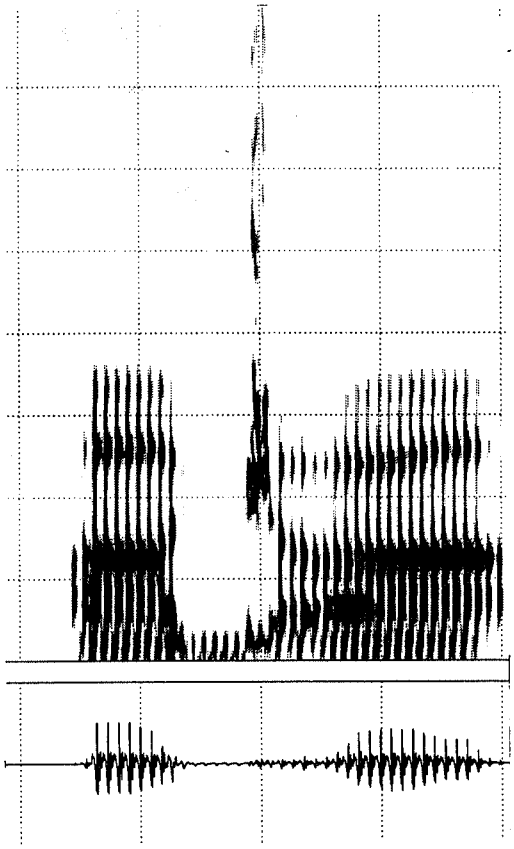


Fig. 4 exemples de règles déclenchées pour la synthèse de [agra]

```
G   ->>  2   AFAMPL := 20
             AVAMPL := 25

G   ->>  1   FORM3 -:= 75 / --- {r}

G   ->>  3   AMPL2 := 50
             AMPL3 := 50
             AMPL4 := 50 / --- {r}

r   ->>  5   AFAMPL := 23
             TRANS1 := 10
             TRANS2 := 10
             F1TRANS1 := 10
             F1TRANS2 := 10 / --- {r} {r} {r}

r   ->>  8   AVAMPL := 28
             AVTRANS1 := 10
             AVTRANS2 := 10
             AFAMPL := 00
             TRANS1 := 10
             TRANS2 := 10
             F1TRANS1 := 10
             F1TRANS2 := 10 / [+vois, +cons] --- {r} {r} {r}
```

19es J.E.P. - Bruxelles, 19 au 22 mai 1992.

SYNTHESE DE L'ARABE STANDARD A PARTIR DU TEXTE PAR TD_PSOLA: LE TRAITEMENT DES PROCESSUS PHONOLOGIQUES

GHAZALI S., ZRIGUI M., BEN MILED Z., JEMNI H.

I.R.S.I.T., TUNIS.

Résumé

A text to speech synthesis system for Arabic is presented in this paper. This system includes two major components: The conversion of Arabic script into phonetic strings, and the synthesis proper; i.e.; algorithms for the generation of speech from phonetic symbols. Arabic graphemes are first converted into phonemes then into allophones following a series of phonological processes of assimilation, juncture, etc...; phonetic strings are then divided into syllables marked for stress placement. Information from the preceding stages are fed into the synthesis program, which includes a dictionary of diphones. The algorithm used for the synthesis is TD_PSOLA applied to diphones concatenation in time domain and adjustment of F0 as well as duration when required. Certain aspects proper to Arabic are discussed such as the production of long vowels and double consonants, the treatment of pharyngalisation and the prosodic modifications for stressed syllables.

I INTRODUCTION

Cet article décrit le système de synthèse de l'arabe standard à partir du texte, en cours de développement à l'IRSIT. Il s'agit d'une synthèse par concaténation de diphones dans le domaine temporel (TD_PSOLA)[1]. L'intérêt de notre approche réside peut-être dans le traitement des processus phonologiques de l'Arabe au niveau de la phonétisation du texte écrit et l'application de ces règles phonologiques au niveau de la synthèse proprement dite.

II QUELQUES CARACTERISTIQUES PHONOLOGIQUES DE L'ARABE STANDARD

L'arabe standard, tel qu'il est prononcé par les locuteurs tunisiens, comprend, au niveau phonétique 27 sons consonantiques (correspondant à 28 graphèmes) et trois voyelles [i, a, u] pouvant chacune être brève ou longue [ii, aa, uu] et ayant chacune deux qualités différentes quand elles sont brèves, selon qu'elles sont en syllabe ouverte ou en syllabe fermée. En outre chaque consonne peut être simple ou géminée. Outre l'importance de la quantité comme trait phonologique pertinent, l'arabe comporte des consonnes pharyngalisées, appelées généralement emphatiques, qui colorent les segments du même mot à des degrés divers [2]. Pour les besoins du traitement, les sons ont été groupés en classes naturelles selon leur comportement phonologique: coarticulation de l'emphase, différents effets co-intrinsèques etc... et un système de notation a été adopté en fonction des contraintes de la machine (Sun 3/150)

III PHONETISATION DE L'ECRITURE ARABE

Puisqu'il s'agit de la synthèse de la parole arabe à partir du texte, nous avons tout d'abord développé un système de transcription graphème-phonème qui tient compte des différents processus phonologiques de l'arabe standard et produit une chaîne phonétique qui constitue l'entrée au synthétiseur. Cette phonétisation du texte écrit comprend plusieurs étapes.

A Transcription des graphèmes

Tout d'abord le texte arabe voyellé est traité et une première chaîne phonétique est obtenue. Trois types de règles de transcription sont appliqués à ce niveau:

- Les règles de transcription des graphèmes arabes en

code interne où un code correspond à chaque graphème.
- Les règles de gémination des consonnes et d'allongement des voyelles.
- Les règles de concaténation des clitiques. Ces règles réajustent la prononciation des mots qui, selon les proclitiques (prépositions+article défini) qui leur sont attachés, peuvent manifester une différence graphique au niveau des frontières des morphèmes sans subir de changement phonétique. Toutes ces règles doivent être appliquées selon un ordre bien déterminé.

A la fin de cette étape le système consulte le dictionnaire d'exceptions, c'est-à-dire, les mots dont la prononciation ne peut pas être déduite des principes généraux décrits par les règles de prononciation.(Par exemple, le mot "celui-ci" est écrit [haðaa] en arabe mais prononcé [haaða] avec la première voyelle longue et la deuxième brève) .

cerner, étant donné que cette langue n'est pas parlée comme langue maternelle: l'influence du dialectal à ce niveau peut engendrer des irrégularités de production. Nos règles préliminaires sont donc modifiées chaque fois que les tests de perception nous l'indiquent.

C La liaison

Il y a un ensemble de règles qui régissent la liaison entre une suite de mots en parole continue en arabe. Cette liaison entraîne une resyllabification au niveau de la jointure des deux mots et peut occasionner une insertion ou un effacement d'un son à cause de la restructuration syllabique. Un autre processus de restructuration opère aussi à la pause: le mot y est prononcé sans marque casuèlle et sans "tanwiin". Les voyelles sont donc effacées et le "tanwiin" est remplacé par la voyelle longue correspondante dans certains cas.

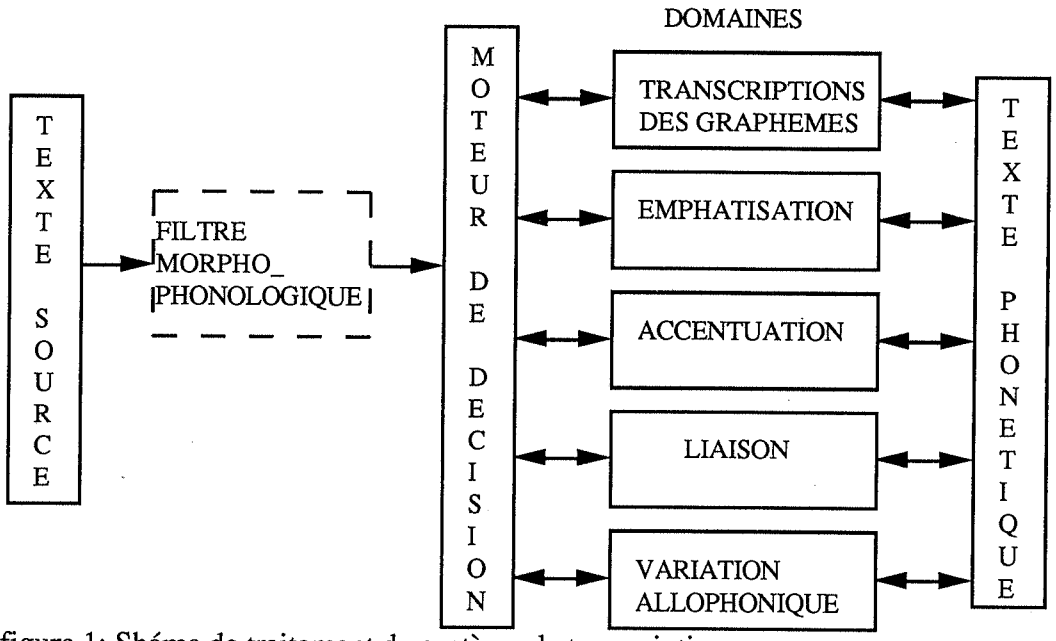


figure 1: Shéma de traitement du système de transcription.

B L'emphatisation et l'accentuation.

Une fois le mot transcrit phonétiquement, le système examine la structure syllabique du mot (nombre de syllabes, poids ou longueur des syllabes) en transformant la chaîne phonétique en une suite de C et V, consulte les règles d'accentuation et place l'accent lexical sur la syllabe appropriée. A ce stade nous ne traitons que l'accent lexical et un seul accent par mot.

Si les règles d'accentuation sont assez simples à implémenter, les règles de propagation de l'emphase en arabe standard sont difficile à

D. Variation allophonique

Il s'agit principalement de règles phonétiques tardives telle que l'assimilation de voisement entre deux obstruentes, ou le relâchement et la centralisation des voyelles brèves en syllabes fermées et d'autres processus articulatoires périphériques naturels ayant un effet sur l'authenticité de parole produite.

Etant donné la quantité d'information importante et variée, l'approche informatique qui convient le mieux pour formaliser la transcription

graphème_phonème est celle d'un système expert. Ce choix se justifie par le fait que dans un système expert les connaissances sont complètement indépendantes des mécanismes de raisonnement. En plus, l'ordre d'activation des régles est guidé par les besoins et la logique du traitement lui même, et donc les mouvements de "va-et-vient" entre les différents niveaux de traitement sont très aisés. Le troisième avantage des systèmes experts est qu'ils permettent une représentation des connaissances complexes sous forme de parcelles élémentaires beaucoup plus faciles à manipuler, telle que les régles de production (formalisme choisi pour représenter les connaissances linguistiques.) [3].

Le shéma général de traitement du système de transcription est illustré par la figure 1. Dans cette figure:

- Le filtre morpho-phonologique permet de corriger dans certains cas les erreurs de voyellation issues de la saisie du texte source. Ce module est en cours de développement [4].
- Le moteur de décision est un processeur intelligent qui permet, selon une certaine logique, d'ordonner l'activation d'une régle d'un domaine, seule, ou en parallèle avec une ou plusieurs autres régles des autres domaines.
- Chaque domaine comprend un certain type de règles (transcriptions des graphèmes,emphatisation, etc ...).

IV DESCRIPTION DU SYSTEME DE SYNTHESE

Le système de synthèse peut être divisé en deux grands modules: le dictionnaire de diphones et la synthèse proprement dite.

A Constitution du dictionnaire:

Le dictionnaire est formé d'un fichier signal en temporel représentant les diphones de l'arabe. Ces diphones sont tirés de logatomes enregistrés par un locuteur tunisien. Les diphones sont représentatifs des contraintes phonétiques de l'arabe. Il s'agit des diphones illustrant les contextes suivants:

1. les groupes consonantiques C1C2 où C1 est différent de C2;
2. la séquence CV où la consonne peut être simple, emphatique ou emphatisée par coarticulation et la voyelle peut être tendue(en syllabe ouverte) ou relâchée (en syllabe fermée);
3. la séquence VC avec les mêmes contraintes phonétiques que la séquence CV;
4. une consonne en fin de mot C# et en début du mot #C;
5. une voyelle en fin de mot (V#). Cette voyelle peut être précédée par une consonne non emphatique, par une consonne emphatique ou emphatisée, par une consonne pharyngale ou par une consonne appartenant au groupe comprenant les uvulaires plus la consonne [r].

Chaque diphone non initial et non final est tiré de la partie centrale non accentuée du logatome. Par exemple le diphone [ta] est choisi dans le logatome [bataba], où l'accent tombe sur la première syllabe.

Les valeurs de l'amplitude et de la fréquence fondamentale sont aussi pris en considération dans le choix des logatomes qui sont toujours compris dans la même phrase porteuse et où la longueur du mot ainsi que les syllabes initiales et finales sont rarement changées.

Une fois que les fichiers parole correspondant aux diphones sont stockés, chaque diphone est traité en marquant la partie stable de deux sons qui le composent ainsi que les périodes des sons voisés. Un exemple de marquage de diphone [ta] est illustré dans la figure 2.

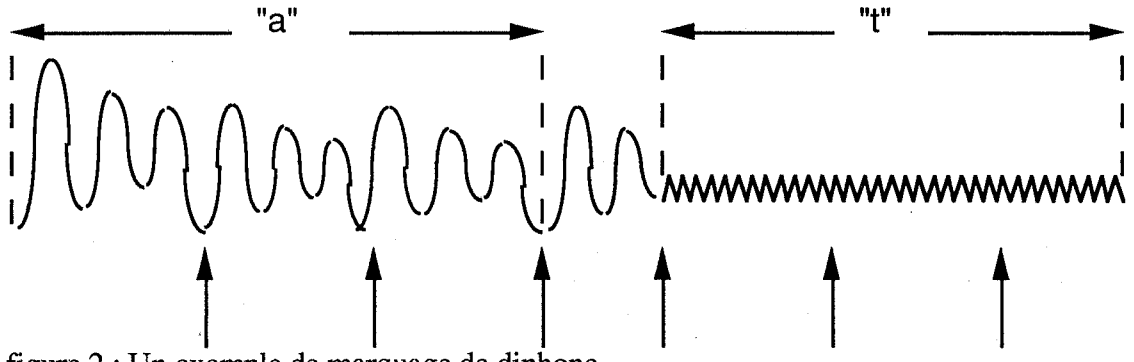


figure 2 : Un exemple de marquage de diphone

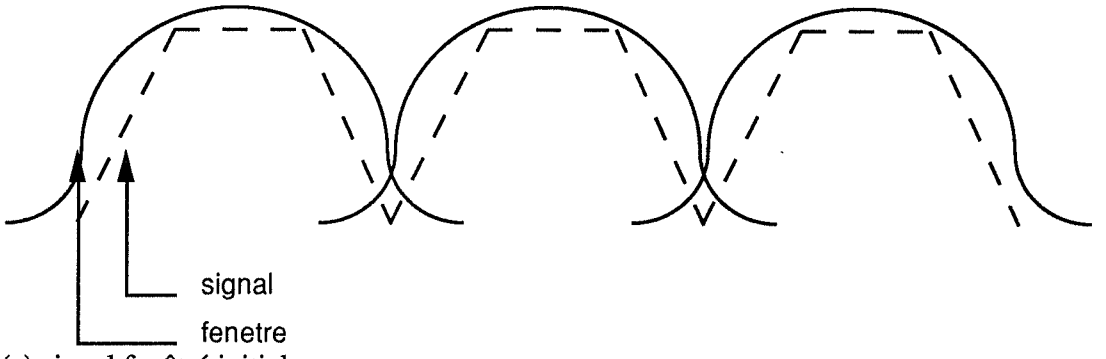
B Synthèse

Le système de synthèse examine la transcription de chaque mot et en effectue le découpage en diphones selon les contraintes indiquées plus haut. Les diphones sont ensuite

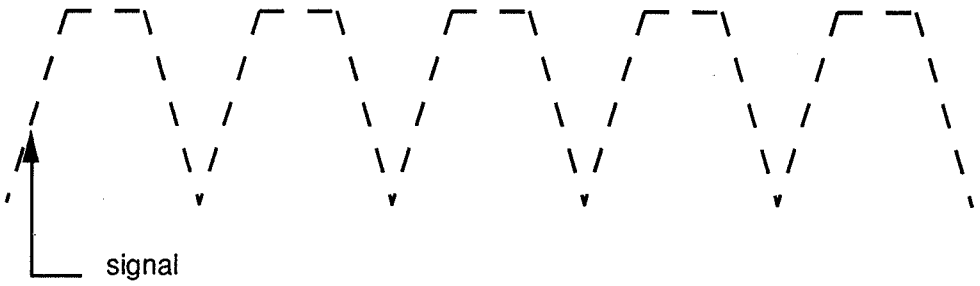
repérés et recherchés dans le dictionnaire. La méthode de synthèse utilisée, TD_PSOLA, consiste à appliquer à chaque marque de F0 d'un diphone une fenêtre de hamming. Chaque partie du signal fenêtré, dénotée par un signal à court terme [5], est ajouté à la partie qui suit s'il y a recouvrement des deux signaux à court terme. On peut donc augmenter ou diminuer la fréquence fondamentale en augmentant ou diminuant la partie de recouvrement de deux parties consécutives. Ceci est illustré dans la figure 3. Par ailleurs deux diphones sont concaténés en recouvrant le dernier signal à court terme du premier diphone avec le premier signal à court terme du second.

incorporant des signaux à court terme entre les signaux à court terme déjà existants. Si le rapport de durée entre un "t" simple et un "t" géminé est de deux, par exemple, chaque signal à court terme est reproduit dans la partie stable "t" des deux diphones. Le même processus est utilisée pour l'allongement des voyelles. Les paramètres de durée vocalique et consonantique ont été calculés sur la base d'une étude portant sur deux locuteurs tunisiens[6].

La correction de la durée consonantique et vocalique est effectuée sur les mots durant la concaténation. Dans l'exemple précédent, si la



(a) signal fenêtré initial



(b) signal après augmentation de FO par un rapport 5/3 et ajustement temporel.

figure 3: Un exemple de changement de FO

1 Ajustement temporel

Dans certains cas, il est nécessaire de modifier la durée d'un son. Ces cas se présentent pour les consonnes géminées et les voyelles longues ou quand il faut corriger la durée intrinsèque ou co-intrinsèque d'une consonne ou d'une voyelle dans un contexte phonétique donné. Par exemple, pour obtenir l'effet de gémination de la consonne [t] dans le mot [kattaba], il faut allonger la partie stable [t] des diphones [at] et [ta].

L'allongement par TD_PSOLA est effectué en ajoutant des périodes ou plus exactement en somme des durées de la partie stable [a] des diphones [ka] et [at] est différente de la durée intrinsèque de la voyelle [a] dans le contexte [at] en syllabe fermée, la durée de la voyelle est corrigée. Ceci est obtenu en ajoutant ou en éliminant des périodes. La même procédure s'applique pour les consonnes.

2. Ajustement de F0

La modification de F0 intervient au niveau de la syllabe accentuée et de la syllabe finale. Cette modification est possible avec TD_PSOLA en

augmentant ou en diminuant la partie de recouvrement de deux signaux à court terme consécutifs. La syllabe accentuée a un F0 supérieur à la syllabe non accentuée. Quant à la syllabe finale, son F0 est inférieur à celui d'une syllabe dans un contexte neutre. A ce stade nous nous limitons à trois modifications de F0 : (a) syllabe accentuée non finale (+ 20Hz), (b) syllabe accentuée et finale (+10Hz), et (c) syllabe finale non accentuée (-20Hz).

La modification de F0 implique bien sûr le changement de la durée totale de la syllabe. Cette modification de la durée du au changement de F0 est compensée par l'ajustement temporel expliqué plus haut..

3 Traitement de l'emphase

Un ensemble de règles régissant la propagation de l'emphase de droite à gauche et de gauche à droite ont été élaborées. Bien que complexes, ces régles ne constituent pas un modèle très fidèle du comportement de ce phénomène de coarticulation, à cause notamment de la nature non binaire de ce trait qui est présent à des degrés divers en fonction de la distance séparant un segment affecté de la consonne emphatique source. Nous avons remarqué , par exemple , que dans un mot de la forme C1V1C2V2....si C1 est emphatique , la synthèse est plus naturelle quant la propagation de l'emphase arrive jusqu'à C2 au lieu de s'arreter à V1. Une voyelle doit être emphatisée totalement (à droite et à gauche), alors qu'une consonne peut jouer le rôle d'un filtre. Une approche phonologique basée sur les unités syllabiques n'est pas d'une grande utilité dans ce contexte [7].

V CONCLUSION

Le système de synthèse actuel génère des mots isolés intelligibles et parfois même assez naturels et agréables à entendre. Son mérite est peut-être de ne pas avoir négligé les spécificités phonologiques de l'arabe. Cependant, il nous reste plusieurs paramètres à mieux maîtriser avant même de passer au contrôle de l'intonation et de la microprosodie. Il s'agit notamment d'une modélisation plus précise de l'organisation temporelle, de l'accentuation et de la restructuration rythmique en parole continue.

Remerciements: Ce projet est mené en collaboration avec le Département Signal de l'ENST (Paris). Nous tenons à les remercier pour leur assistance notamment dans la maîtrise de TD_PSOLA.

BIBLIOGRAPHIE

[1].E. Moulines, "Algorithmes de codage et modification des paramètres prosodiques pour la synthèse de la parole à partir du texte", Thèse de doctorat, Ecole Nationale Supérieure des Télécommunications Paris, 1990.

[2] S. Ghazali, "Back consonants and backing coarticulation in Arabic", Thèse de Ph.D., Université de Texas, 1977.

[3] M. Zrigui, " Elaboration d'un nouveau système braille arabe basé sur la phonétique", Thèse de doctorat, Université Paul Sabatier Toulouse, 1987.

[4] J.P. Haton, "Utilisation des techniques à bases de connaissances en reconnaissance automatique de la parole", Rapport interne, CRIN, 1985.

[5] E. Moulines & F. Charpentier, "Pitch-synchronous waveform processing techniques for texte-to-speech synthesis using diphones",Speech Communication,1990.

[6] S. Ghazali & A. Braham, " Voyelles longues et voyelles brèves en arabe standard: organisation temporelle ", JEP, 1990.

[7] S. Ghazali, "La diffusion de l'emphase: l'inadéquation d'une solution tauto-syllabique", Analyse Théorie, 1981.

19es J.E.P. - Bruxelles, 19 au 22 mai 1992.

UNE APPROCHE « ORIENTÉE LEXIQUES » POUR LA GÉNÉRATION AUTOMATIQUE DE L'INTONATION

Véronique Aubergé

INSTITUT DE LA COMMUNICATION PARLÉE
URA CNRS n° 368 INPG/ENSERG Université Stendhal BP 25
38040 Grenoble cedex 9 France

Résumé

Chaque situation de synthèse vocale nécessite une stratégie prosodique spécifique. Nous présentons ici une méthodologie et des outils pour la constitution semi-automatique d'un module de génération de l'intonation dans un système de synthèse.

Ce travail consiste d'abord en l'analyse d'un corpus enregistré par un locuteur de référence. La clef de voûte de ce corpus est la notion de *rendez-vous* à différents niveaux des structures linguistiques (la phrase, la proposition, le groupe et le sous-groupe) entre des *unités globales* du texte et de l'intonation. Pour chacun de ces niveaux, les contours sont analysés et rassemblés dans des classes, chacune d'elles étant identifiée par une liste spécifique d'attributs. La cohérence, comme la disjonction des classes, sont vérifiées selon un principe de paires minimales sur les attributs. Un *contour-moyen* statistiquement représentatif est ensuite calculé pour chaque classe.

La formalisation issue de l'analyse est un *lexique dynamique hiérarchisé de formes intonatives globales* (les contours-moyens). L'activation d'une entrée par appariement de la liste d'attributs entraîne le calcul inclusif, dans un ordre *top-down*, de l'intonation d'une phrase.

1. INTRODUCTION

Le traitement traditionnellement modulaire de l'intonation dans les systèmes de synthèse à partir du texte (SSpT) pourrait laisser croire à l'existence linguistique d'une intonation neutre (plate), où les informations véhiculées par les structures intonatives seraient simplement absentes. En fait, il est clair qu'une intonation "mal formée" induit une communication anormale, et donc un processus brouillé de décodage : l'intonation est une composante inhérente à la langue orale [Bolinger, 1989] et une situation "sans prosodie" ne peut pas être une situation réaliste pour un SSpT.

La description des structures prosodiques de la parole naturelle fait référence aux faits linguistiques et extra-linguistiques qui caractérisent le message [Grosjean, 1983]. Dans un SSpT, les structures prosodiques sont générées à partir des seules données textuelles qui ne contiennent aucune trace du modèle du locuteur. Le module de génération de l'intonation doit donc intégrer, explicitement ou implicitement, un modèle du locuteur.

La solution que nous proposons ici fait implicitement référence à un modèle du locuteur, puisqu'elle est issue d'une méthodologie inductive. La première étape de ce travail est la constitution d'un corpus à partir d'un ensemble de paires minimales d'attributs. Ils caractérisent, aux niveaux successifs des structures linguistiques que sont la phrase, la proposition, le groupe et le sous-groupe, les points de *rendez-vous* entre des *unités globales* du texte et de l'intonation. Pour chacun de ces niveaux, des listes spécifiques d'attributs définissent des classes de contours extraits du corpus. La cohérence des contours à l'intérieur d'une même classe, comme la disjonction de deux classes, sont assurées selon un principe de paires minimales sur les attributs. L'unité minimale de description d'un contour est la syllabe. La moyenne des contours de chaque classe, après que sa validité statistique soit vérifiée, est représentée par un *contour-moyen* associé à sa liste d'attributs.

Ces *formes intonatives globales* sont conservées dans un *lexique* organisé par niveaux. Le calcul d'un patron intonatif est décomposé en deux phases : l'une suprasyllabique, l'autre subsyllabique. La première étape consiste, niveau par niveau, depuis la phrase jusqu'au sous-groupe, à activer les formes du lexique appariées avec les attributs, puis à inclure le patron du niveau courant à l'intérieur du niveau précédent. Le patron subsyllabique est ensuite calculé à partir du patron suprasyllabique et des contours segmentaux associés aux niveaux terminaux du lexique.

2. LE CORPUS

2.1. Les hypothèses sous-jacentes

La structure du corpus doit rendre compte des hypothèses qui sous-tendent notre démarche : les attributs présupposés comme points de rendez-vous sont représentés dans le corpus pour un sous-ensemble vaste de leur domaine de variation, et ceci sont la contrainte de paires minimales. Ces attributs ont été choisis en fonction de critères phonotactiques, syntactiques, ou sémantiques [Rossi *et al., 1981* ; Grosjean & Dommergues, 1983 ; Hirst, à paraître]. Nous nous sommes restreints à une situation de lecture, par un seul locuteur, de phrases isolées, minimisant ainsi les relations aux faits linguistiques tels que l'organisation du discours ou du dialogue [House *et al.*, 1990].

Des formes linguistiques ont été définies pour la phrase, la proposition, le groupe, et le sous-goupe (composant des groupes "longs", *i.e.* supérieurs à quatre syllabes). Une hypothèse forte de cette étude est que des formes intonatives sont attachées aux mêmes points d'articulation que les formes linguistiques, et surtout que ces formes seront considérées comme des unités globales du traitement intonatif.

Les paramètres prosodiques choisis pour la définition des contours sont la fréquence fondamentale (Fo) et la durée. L'importance relative, voire la dépendance, de ces deux paramètres étant variable selon le choix stratégique du locuteur [Caelen, 1991], nous avons arbitrairement choisi Fo comme paramètre directeur classificateur dans l'analyse du corpus, en raison de sa plus grande facilité interprétative.

La durée est traitée en deux étapes. Au niveau subsyllabique, un paramètre temporel est codé pour chaque son par une valeur de début et de fin. Ensuite, une *durée syllabique* (Ds) est calculée pour chaque syllabe, par soustraction de valeurs intrinsèques référentielles de chaque son de la syllabe (modulo des coefficients de coarticulation). Le paramètre Fo est codé par trois valeurs : début, maximum et fin des noyaux vocaliques [Émerard et Benoît, 1989].

2.2. Les paires minimales

L'établissement factoriel du corpus respecte un principe de paires minimales sur les valeurs des attributs : pour toute liste d'attributs instanciés, on trouve obligatoirement la même liste d'attributs qui varie seulement d'une instance. La comparaison des contours, caractérisés par ce couple de listes d'attributs, confirmera ou non la paire minimale. La combinatoire sur ces paires minimales doit respecter des contraintes statistiques afin que le cardinal d'une classe de contours puisse éventuellement valider la moyenne sur la classe.

Le nombre de syllabes de l'unité est un attribut récurrent à chaque niveau considéré : cet attribut de longueur est le relai symbolique de la durée. Les autres attributs sont définis spécifiquement à chaque niveau.

Le corpus totalise 164 phrases. Le locuteur (âgé de 40 ans environ et originaire de l'Ile de France) a également enregistré le corpus dont sont extraits les polysons utilisés [Aubergé, 1991] dans notre système de synthèse.

2.2.1. Le niveau phrase

Le principal attribut de ce niveau (cf les exemples de Tableau 1) est la modalité qui prend les valeurs : déclarative (positive *vs.* négative), interrogative (introduite *vs.* inversée *vs.* elliptique) et impérative. La longueur varie de 3 à 21 syllabes.

Tableau 1 :

attributs		*exemples*
déclarative	positive	Je peux passer.
	négative	Je *ne* peux *pas* passer.
interrogative	directe	Je peux passer ?
	introduite	*Est-ce*-que je peux passer ?
	inversée	*Puis-je* passer ?
impérative		Passez !

2.2.2. Le niveau proposition

Les attributs de ce niveau caractérisent principalement la nature des relations de dépendance de chaque proposition, sa position dans la phrase et sa nature (voir les exemples donnés dans le Tableau 2). La longueur des propositions varient entre 3 et 9 syllabes.

Tableau 2 :

attributs				*exemples*
position absolue	init.			P 1
	finale			P 2
indépen.	isolée			P 3
	juxta-posée			P 4
dépend.	verbe	dominante		P 5
		dominée		P 5'
	nom	dominante		P 6
		dominée	relative	P 6'
			insérée	P 7

P1 : *Quand l'enfant pleurait*, il était malade.
P2 : Il était malade *quand l'enfant pleurait.*
P3 : *Ils jouent avec un balai.*
P4 : Je vois ces enfants ; *ils jouent avec un balai.*
P5 : *Je vois ces enfants* quand ils jouent avec un balai.
P5' : Je vois ces enfants *quand ils jouent avec un balai.*
P6 : *Je vois ces enfants* qui jouent avec un balai.
P6' : Je vois ces enfants *qui jouent avec un balai.*
P7 : Ces enfants, *ils jouent avec un balai*, je les vois.

2.2.3. Le niveau groupe

Le groupe est défini ici comme le constituant syntaxique inférieur à la proposition. Les attributs définis pour le groupe sont multiples (cf. exemples dans le Tableau 3) : la nature (groupe nominal, verbal – au sens restreint – adjectival, adverbial,

grammatical), la fonction, la position relative (fonction d'autres groupes) et absolue (initiale, interne ou finale de proposition). Le groupe nominal (GN) a plus particulièrement été représenté dans le corpus, puisqu'il est potentiellement le plus complexe dans ses constructions.

La longueur d'un groupe varie de 2 à 15 syllabes. Lorsqu'un groupe est supérieur à quatre syllabes alors il est décomposé en sous-groupes. Le groupe de quatre syllabes ou moins est un niveau terminal du domaine suprasyllabique.

Tableau 3 :

attributs		*exemples*
nature	groupe nominal	P 8
	groupe verbal	P 8'
function (GN)	sujet	P 9
	objet	P 10
position absolue	initiale	P 11
	interne	P 12 13
	finale	P 14
position relative (GN)	pré-verbale	P 9"
	post-verbale	P 15

P8 : *Ce passant* est passé.
P8' : Ce passant *est passé*.
P9 : *Ce passant* chantait.
P10: Je vois *ce passant*.
P11: *Souvent*, un passant chantait l'opéra.
P12: Un passant, *souvent*, chantait l'opéra.
P13: Un passant chantait *souvent* l'opéra.
P14: Un passant chantait l'opéra *souvent*.
P9' : *Ce passant* chantait.
P15: On entendait *ce passant*

2.2.4. Le niveau sous-groupe

Le sous-groupe est défini comme constituant inférieur aux groupes longs et récursivement aux sous-groupes longs. Il varie de 2 à 12 syllabes. Les attributs du sous-groupe (cf exemples dans le Tableau 4) sont la valeur catégorielle (nom, adverbe – autonome vs. modifieur - mot grammatical, verbe - outil, composé, conjugué, infinif - adjectif), la position relative (adjectif / nom, nom / verbe) et absolue (initiale, interne ou finale de groupes).

Tableau 4 :

attributs			*exemples*
groupe simple	nature	nom	P 16
		adjectif	P 16'
	position relative	pré-nominale	P 16"
		post-nominale	P 17
groupe complexe	position d'énumeration	initiale	P 18
		interne	P 19
	dépendance	GN	P 20

P16 : ce fantastique *passant*
P16' : ce *fantastique* passant
P16": ce *fanstastique* passant
P17 : ce passant *fantastique*
P18 : *du vin*, du pain, du boursin
P19 : du pain, *du vin*, du boursin
P20 : le pas *de ce fantastique passant*

3. LA MÉTHODE D'ANALYSE

Le corpus est l'association des codages phonétiques, des étiquettes linguistiques (les attributs) et des codages subsyllabiques des paramètres prosodiques. Un ensemble de gestionnaires permet l'analyse des données symboliques et des codages physiques.

Rappelons que l'hypothèse principale est l'association d'un contour intonatif global à un ensemble d'attributs. Les contours sont analysés d'abord dans le domaine suprasyllabique. Aux niveaux non terminaux (la phrase, la proposition et le groupe "long"), les contours sont caractérisés par une *ligne de déclinaison* - définie ici par le Fo syllabique de la première et de la dernière syllabe de la phrase, proposition ou groupe - et par le nombre de syllabes de l'unité. Aux niveaux terminaux (le groupe "court" et le sous-groupe), les contours sont définis par Ds et Fo syllabique de chaque syllabe successive de l'unité. À chaque contour suprasyllabique d'un niveau terminal, est associé un contour subsyllabique (codage de Fo en trois points par voyelle).

À chaque niveau, les contours ont été automatiquement segmentés et regroupés selon les valeurs des attributs instanciés au niveau considéré. L'homogénéité de chaque classe a été vérifiée visuellement - à l'aide d'outil de superposition - puis objectivement : la moyenne des contours ou *contour-moyen* (CM) de chaque classe a été systématiquement calculée (cf. Figure 1). Les paires minimales sur les CM (et aussi sur les contours originaux) sont systématiquement testées : losque la paire n'est pas contrastive, les deux classes sont unifiées.

L'ensemble des CM, chacun étant associé à sa liste d'attributs constitue un *lexique* hiérarchique de formes intonatives indexé par les attributs [Aubergé, 1991].

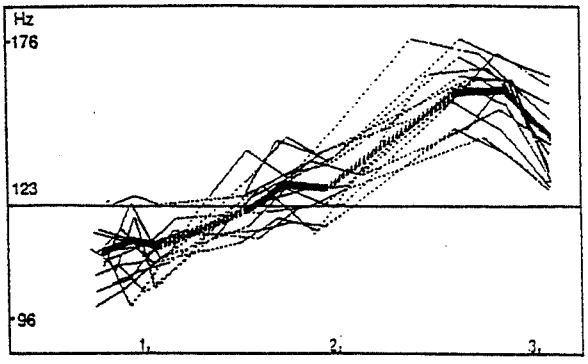


Figure 1 . Les GN de 3 syllabes, en fonction non-complément, en position préverbale et début de proposition (le CM est en gras)

L'analyse du corpus a confirmé pour l'essentiel les hyopthèse faites *a priori* [Aubergé, 1992]. Il en ressort que les contours non-terminaux pourraient sans doute être caractérisés par des formes plus fines qu'une simple ligne de déclinaison.

4. LA GÉNÉRATION DE L'INTONATION

4.1. La méthode

L'entrée textuelle du SSpT est traitée phrase par phrase. Une analyse morpho-syntaxique délivre une partie des attributs nécessaires à l'indexation du lexique [Rouault, 1988]. Après la phonétisation et la syllabation de la phrase [Aubergé, 1991], le patron intonatif est calculé en étapes successives. Un patron intonatif suprasyllabique est calculé par phases successives : pour chaque niveau, l'un après l'autre, l'un en fonction de l'autre, et dans un ordre *top-down* de la phrase vers le groupe ou le sous-groupe. À chaque niveau, le patron de Fo est extrait du lexique par appariement des attributs. Ce patron est ensuite aligné sur le patron du niveau supérieur. La procédure d'alignement est illustrée par l'exemple qui suit.

Par la suite, ce patron suprasyllabique est transformé en un patron sub-syllabique. À chaque son est associée une durée, en ajoutant les valeurs intrinsèques modifiées par les coefficients cointrinsèques des sons à Ds. Le codage phonétique de Fo en trois points pour chaque voyelle est interpolé sur la chaîne acoustique par une fonction Spline cubique. La figure 2 décrit la procédure de calcul de l'intonation.

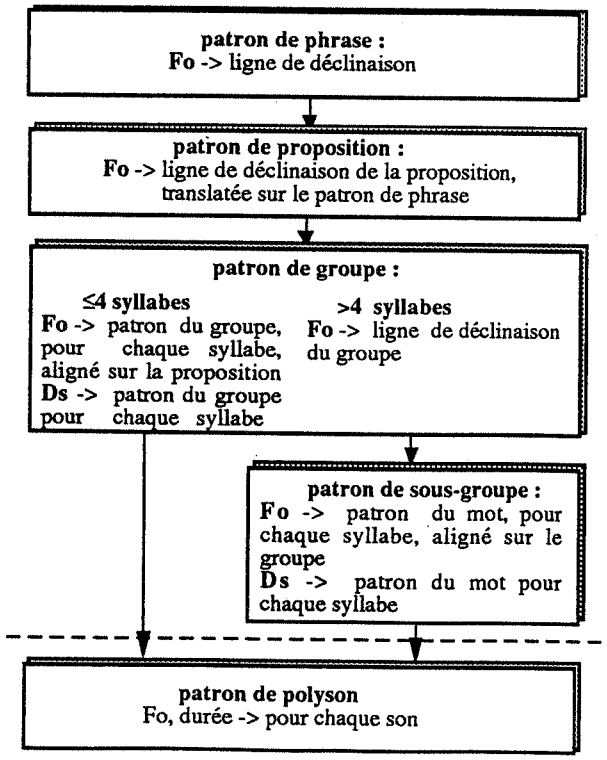


Figure 2 . Le calcul hiérarchique de l'intonation

4.2. Un exemple

Nous décrivons ici le calcul de la phrase :

Tous les six mois, ce passant chantait l'opéra.
[tu le si mwɑ sə pasɑ̃ ʃɑ̃tɛ lɔpeʀa]

4.2.1. Le calcul suprasyllabique

Les attributs des différents niveaux étant calculés pour la phrase à générer, le calcul commence par l'accès au CM du niveau phrase, comme le décrit le Tableau 5.

Tableau 5 :

attributs	phrase
long (syllables)	12
modalité	déclarative(assertive)

Les attributs du niveau proposition n'indexent pas de CM du lexique, puisque la proposition est égale à la phrase.

Les attributs du niveau groupe sont appariés avec les index du lexique selon le Tableau 6.

Tableau 6 :

attributs	*groupe 1* (tu- le- si- mwɑ)	*groupe 2* (sə- pa-sɑ̃)
long (syl.)	4	3
valeur	GN	GN
fonction	complément	non-complément
pos. abs.	initiale	interne
pos. rel.	pré-verbale	pré-verbale

attributs	*groupe 3* (ʃɑ̃-tɛ)	*groupe 4* (lɔ-pe-ʀa)
long (syl.)	2	3
valeur	groupe vebal	GN
fonction	conjugué	non-complément
pos. abs.	interne	finale
pos. rel.		post-verbale

Les CM extraits pour le niveau groupe (Patron 2) sont alignés sur le Patron 1. Le résultat de l'alignement est le Patron 3 (cf. Tableau 7 et Figure 3)

Tableau 7 :

nb syl.	1	2	3	4	5	6
Patron 1	108	106	104	101	99	97
Patron 2	118	109	116	137	108	119
Decl.2	118	114	111	108	104	101
différence relative	-10	-8	-7	-6	-5	-3
Patron 3	108	101	109	130	103	115

nb syl.	7	8	9	10	11	12
Patron 1	95	93	91	89	87	85
Patron 2	146	110	119	108	99	81
Decl.2	97	94	91	87	84	81
différence relative	-2	-1	0	+1	+3	+4
Patron 3	144	109	119	109	102	85

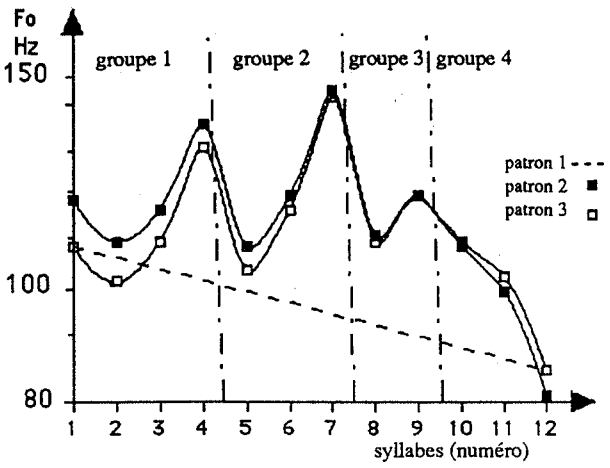


Figure 3 . Alignement du Patron 2 (CM niveau groupe) sur Patron 1 (CM niveau phrase), le résultat est dans Patron 3

Nous montrons ici seulement le calcul du paramètre directeur Fo, le calcul de la durée est décrit dans [Aubergé, 1991].

4.2.2. Le calcul suprasyllabique

Les CM activés par les attributs du Tableau 7 sont des CM terminaux, puisqu'ils sont tous de longueur inférieure ou égale à 4. On passe donc à la deuxième phase du calcul qui est la détermination du patron subsyllabique, par combinaison des trois valeurs codées de Fo (cf. Tableau 8).

Tableau 8 :

nb syl.	1	2	3	4	5	6
1e val Fo	+2	-2	-1	-12	0	-4
2e val Fo	+3	+2	+3	-1	+1	0
3e val Fo	-6	-2	-4	+11	-1	0

nb syl.	7	8	9	10	11	12
1e val Fo	+3	+2	-4	+2	+5	+3
2e val Fo	+6	0	+4	-1	-1	-1
3e val Fo	-8	-3	+1	-2	-3	-2

CONCLUSION

Il nous semble intéressant de retenir que la double hypothèse sous-jacente à cette étude - (1) l'existence de formes intonatives caractéristiques, autrement dit d'unités de traitement, des niveaux phrases, proposition, groupe et sous-groupe (2) la dépendance hiérarchique de ces unités - semble bien être vérifiée par l'analyse du corpus : ces formes globales ont pu être calculées avec les contours-moyens. Peut-être pourrait-on trouver une corroboration à ces résultats dans une étude psycholinguistique des manifestations de l'intonation : pourrait-on considérer que le calcul mental de l'intonation procède par un accès lexical à un lexique de formes intonatives, organisé par niveau et dont l'accès serait direct à chaque niveau. Autrement dit, des formes intonatives correspondant à chacun des niveaux pourraient-elles être des unités du traitement cognitif de l'intonation [Grant *et al.*, 1985] ?

L'inconvénient majeur de tout lexique est bien entendu qu'il ne décrit pas un domaine exhaustif. Ce lexique cependant est particulier, puisqu'il intégre une structure hiérarchique et, pour chaque entrée, un ensemble de paramètres qui le transforme, en quelque sorte, en un *lexique fonctionnel*. C'est pourquoi il nous semble que la généralisation de cette étude peut être menée selon deux voies disjointes. Le premier type d'extension que nous imaginons est somme toute dans la même logique : il s'agirait de généraliser le lexique par un modèle stochastique de type réseau connexionniste, ainsi l'unification de deux classes, fonction subjective dans cette méthodologie, serait remplacée par une procédure objective de calcul implicite. Des applications déjà réalisées dans d'autres études [Traber, 1990], semblent particulièrement prometteuses, et ce serait sans doute une solution intéressante pour la synthèse. Une autre démarche possible serait d'expliciter le niveau phonologique sous-jacent au lexique (représenté par le statut "fonctionnel" du lexique), et de déboucher peut-être, par le jeu des unifications sur les classes, sur un modèle. Ce modèle serait alors dépendant des décisions de l'expert qui le formaliserait.

Dans la première application que nous avons faite de cette méthodologie, le corpus représente principalement les variations syntaxiques du GN, sans doute faudrait-il d'une part l'étendre aux autres structures syntaxiques et mettre en évidence les poids sémantiques relatifs à certains attributs syntaxiques (nous avons pu le constater sur l'exemple des conditionnelles, et des adverbes), et d'autre part dépasser le niveau de la phrase, en ajoutant les attributs susceptibles de caractériser le niveau du discours. Cette application s'est intéressée à un corpus lu, et bien que l'étude de l'intonation dans le dialogue en soit à ses débuts, il est envisageable de définir selon la même méthodologie un corpus représentant une structure restreinte de dialogue qui pourrait correspondre à un dialogue homme-machine. D'une manière générale, il faudra choisir un contexte et une ou plusieurs stratégies de l'intonation pour une application ciblée de la synthèse [House & Youd, 1990], et recommencer l'analyse d'un corpus spécifique.

Dans l'analyse que nous avons faite de l'intonation du corpus, le paramètre classificateur a été uniquement Fo, la durée n'intervenant que par le biais du nombre de syllabes, et Ds n'étant attribuée qu'au niveau terminal. Ce choix a été principalement guidé par une contrainte : il n'existe pas, à notre connaissance,

d'études qui mettent en évidence des unités temporelles dont la variation coïncident clairement avec une variation intonative. Il faudrait pouvoir distinguer ce qui, dans la durée, relève du niveau segmental et suprasegmental, et disposer de plus d'un modèle qui mette en évidence la corrélation entre les différents paramètres acoustiques. A ce stade est posé à nouveau le problème de l'indentification acoustique de la syllabe, et en amont de son existence phonétique, voire phonologique. Les contours de durée syllabique n'ont pas été étudiés dans cette étude, mais le choix que nous avons fait de décomposer l'analyse de la durée en une phase syllabique puis segmentale, nous laisse envisager une continuation possible de ces travaux dans cette direction [Campbell, 1989].

RÉFÉRENCES

Aubergé V., "Developing a Structured Lexicon for Synthesis of Prosody", *In Talking Machines*, Bailly G. & Benoît C. eds., North Holland, 307-322, 1992.

Aubergé V., *La synthèse de la parole, des règles aux lexiques* , Thèse de doctorat, Univ. P. Mendès France, Grenoble, 210 p., 1991.

Bartkova K. & Sorin C., "A model of segmental duration for speech synthesis in French," *Speech Communication*, 6, 245-260, 1987.

Bolinger D.L., *Intonation and its uses*, Arnold ed., Hodder & Stoughton, 1989.

Caelen G., *Stratégies des locuteurs et consignes de lecture d'un texte : analyse des interactions entre modèles syntaxiques, sémantiques, pragmatique et paramètres prosodiques*, Thèse d'Etat, Aix-en-Provence, 1991.

Campbell W. N., "Syllable-level duration determination." *In J.P. Tubach & J.J. Mariani (Eds.), Proceedings of the European Conference on Speech Communication and Technology (Eurospeech 89, Paris, September 1989)*, 698-701, CEP Consultants Ltd., Edinburgh, 1989.

Émerard F. & Benoît C., "De la production à l'extraction, l'état d'un chantier," *16èmes Journées d'Études sur la Parole*, 224-228, 1987.

Grant S.R. & Dingwall W. O., "The Role of the Right Hemisphere in Processing Linguistic Prosody", *13th An. Meet. of the INS*, San Diego, 1986.

Grosjean F., "How long is the sentence? Prediction and prosody in the on-line processing of language," *Linguistics*, 21, 501-529, 1983.

Grosjean F. & Dommergues J.Y., "Les structures de performance en psycholinguistique," *L'Année Psychologique*, 83, 513-536, 1983.

House J. & Youd N., "Contextually appropriate intonation in speech synthesis," *Proceedings of the ESCA Workshop on Speech Synthesis, Autrans, France.*, 1990.

Hirst D. (in press), "Detaching intonational phrases from syntactic structure," *Linguistic Inquiry.*

Rossi M., Di Cristo A., Hirst D., Martin P., & Nishinuma Y., *L'intonation, de l'acoustique à la sémantique*, Klincksieck Ed., Paris, 1981.

Rouault J., *Linguistique automatique, Applications documentaires*, Sciences pour la Communication, Peter Lang, Berne, 309 p, 1988.

19es J.E.P. - Bruxelles, 19 au 22 mai 1992.

REALITE PSYCHOLOGIQUE DES UNITES PHONOLOGIQUES ET SYSTEMES D'ECRITURE

DANIEL HOLENDER

LABORATOIRE DE PSYCHOLOGIE EXPERIMENTALE
UNIVERSITE LIBRE DE BRUXELLES

Parmi les défis auxquels les disciplines cognitives se trouvent confrontées, la question de savoir comment s'effectue la médiation entre le monde physique et les représentations mentales conscientes est certainement l'une des plus difficiles à résoudre. La réponse principale fournie par la psychologie cognitive contemporaine consiste à postuler qu'entre les mécanismes de transduction neurophysiologiques et les représentations mentales accessibles à l'introspection, il existerait des processus mentaux inconscients qui extraient et traitent l'information contenue dans les objets et les événements du monde physique. Ces processus opéreraient sur des représentations mentales elles aussi inaccessibles à la conscience.

La version que Fodor (1983) a donnée de l'approche cognitive contemporaine reflète assez bien le consensus qui s'était progressivement établi depuis les années cinquante, parfois de manière tacite et peu explicite, chez la plupart des psychologues expérimentaux. La notion fondamentale est qu'une partie de l'esprit est organisée de manière modulaire. Un module est un dispositif de traitement de l'information qui possède sa propre base de données et qui est informationnellement cloisonné. Ceci signifie que le traitement s'effectue en vase clos sans qu'un module ait accès à l'information contenue dans d'autres modules ou à celle contenue dans les processus cognitifs non modulaires de la pensée qui constituent la vie mentale consciente. Un module est aussi impénétrable cognitivement, c'est-à-dire que ni les représentations symboliques qu'il contient, ni les opérations de traitement qu'il effectue ne sont accessibles à la conscience.

Quels sont les processus mentaux dont on peut penser qu'ils ont une organisation modulaire? Pour Fodor, les meilleurs candidats sont certainement les systèmes d'entrée évoqués plus haut; les systèmes qui effectuent la médiation entre les processus physiologiques de transduction et les processus cognitifs accessibles à l'introspection. En outre, certains domaines spécialisés du traitement de l'information, comme la perception de l'espace et le langage, ou au moins certaines de ses composantes, sont également considérés comme de bons candidats pour ce type d'organisation. Par exemple, Liberman et Mattingly (1985, 1991) proposent qu'un module unique est responsable à la fois de la perception et de la production de la parole. Sans être totalement identique à celle de Fodor, la position de Chomsky (Chomsky, 1980, par exemple) sur l'existence de représentations symboliques et de règles inconscientes correspondant à différentes composantes du langage se situe clairement dans la même perspective. Notons toutefois que l'existence de représentations mentales symboliques en principe inaccessibles à la conscience est quand même parfois mise en doute, notamment par Searle (1990).

Historiquement, la question de la réalité psychologique des entités postulées par les linguistes s'est d'abord posée par rapport aux représentations mentales conscientes. Avec l'avènement de l'approche cognitive contemporaine et la croyance en l'existence de représentations mentales symboliques inaccessibles à la conscience, la même question peut maintenant être posée à ce niveau également. Une fois cette dichotomie des processus mentaux admise, le problème des relations entre entités conscientes et inconscientes émerge tout naturellement.

1. LES DEUX FORMES DE REALITE PSYCHOLOGIQUE DES CONCEPTS LINGUISTIQUES.

Pour rendre compte de la manière dont la parole peut être associée à du sens, les linguistes sont amenés à postuler divers constructs hypothétiques. Par exemple, les notions de trait phonétique, de phonème, de segment, de syllabe, de constituant syllabique, de more, de morphème sont autant de constructs qui peuvent entrer en jeu dans la description des processus phonologiques. Comme on l'a déjà signalé, la question de la réalité psychologique de tels constructs a d'abord été posée par

rapport aux représentations mentales conscientes. Dans l'esprit du linguiste, les entités qu'il invente et manipule sont ipso facto psychologiquement réelles puisqu'elles correspondent à des contenus de conscience objectivables pour lui. La question est moins triviale quand il s'agit de déterminer lesquelles parmi ces entités correspondent le mieux aux intuitions du locuteur naïf.

Sapir (1933), s'interrogeant sur les préférences d'informateurs amérindiens quant à la transcription phonétique ou phonémique de leur langue non écrite, concluait en faveur de la seconde possibilité, croyant ainsi démontrer le primat de l'intuition phonémique sur l'intuition phonétique. Mais, contrairement à ce que pensait Sapir, ses informateurs étaient loin d'être naïfs du point de vue de l'analyse segmentale de la chaîne parlée. En effet, sachant tous lire et écrire l'anglais, ils étaient tous en possession du code alphabétique, donc d'une représentation phonémique du langage. Or, il est maintenant bien établi que la maîtrise d'une écriture alphabétique nécessite un degré élevé de conscience phonémique et, qu'en retour, l'apprentissage de la lecture dans une telle écriture promeut cette prise de conscience (voir, par exemple, Morais, Alegria, & Content, 1987). En outre, la connaissance d'un alphabet implique que la notion de segment phonémique est lexicalisée puisque chaque nom de lettre réfère à un phonème. Or, le fait qu'un concept soit lexicalisé a des conséquences non négligeables sur la pensée et sur la pragmatique de la communication.

En effet, bien que dans sa formulation la plus stricte, le relativisme linguistique -l'idée que le langage détermine la pensée et la perception- soit généralement considéré comme insoutenable, des formes plus faibles de l'hypothèse whorfienne (Whorf, 1956) comme, par exemple, celle récemment défendue par Hunt et Agneli (1991), sont éclairantes pour la discussion en cours. Selon ces auteurs, la question doit être posée en termes du degré avec lequel certaines manières de penser le monde sont plus naturelles que d'autres en fonction des propriétés de la langue qu'on parle. Il est, par exemple, plus aisé et plus naturel de penser et de communiquer un certain concept si celui-ci possède un nom que dans le cas contraire. La lexicalisation d'un concept, en focalisant l'attention sur certains aspects du monde, peut aussi parfois biaiser la description de la réalité tant il est commode de communiquer en utilisant les termes lexicaux qui sont les éléments primitifs du discours plutôt que de longues descriptions détaillées.

Un autre point important est que les catégories naturelles ne sont pas aristotéliciennes, c'est-à-dire que l'appartenance à une catégorie n'est pas définie par des conditions nécessaires et suffisantes mais de manière probabiliste (Rosch, 1973). Ceci résulte presque nécessairement du fait que la plupart des concepts exprimables par le langage repose sur un découpage discret d'une réalité qui se présente presque toujours de manière continue. Dans ces conditions, ils est inévitable que les limites des concepts soient relativement indéterminées (floues) et que certains spécimens d'une classe conceptuelle soient considérés comme plus représentatifs que d'autres.

Bien entendu, le point précédent s'applique à tous les concepts, qu'ils soient lexicalisés ou non. Envisageons toutefois un scénario plausible de l'impact de la lexicalisation sur la classification d'événements ou d'objets. S'il s'agit de spécimens typiques non ambigus, il y a gros à parier que leur assignation à un concept déterminé sera la même que le concept ait un nom ou qu'il doive être défini par la liste de ses propriétés. En revanche, comme on l'a déjà souligné plus haut, on peut s'attendre à ce que des représentants ambigus soient beaucoup plus fréquemment assignés à un certain concept quand celui-ci a un nom que dans le cas contraire. En effet, la nécessité d'utiliser une description plutôt qu'une désignation pourrait amener à être plus précis et plus sélectif quant aux propriétés jugées pertinentes, ce qui pourrait parfois se traduire par la définition d'une nouvelle catégorie ou sous-catégorie conceptuelle.

Enfin, un objet quelconque peut généralement être classé à plusieurs niveaux différents dans une hiérarchie fondée sur la relation d'inclusion de classes. Or, il est maintenant bien établi que les différents niveaux d'une telle hiérarchie ne sont pas équivalents. Il semble qu'il existe un niveau fondamental de catégorisation qui est psychologiquement plus naturel que les autres. Une des propriétés de ce niveau est d'être celui auquel les objets sont le plus spontanément nommés. Les travaux de Rosch et ceux qu'elle a inspirés (Cf. Mervis & Rosch, 1981, pour une revue) ont montré que ce niveau dépend partiellement de propriétés perceptives mais est également influencé par des facteurs linguistiques et sociologiques ainsi que par le degré d'expertise. Par exemple, pour la plupart d'entre nous, il existe une grande variété de petits animaux qui sont simplement qualifiés d'insectes et dont très peu peuvent être identifiés par leur nom. Le nombre d'insectes nommables et la possibilité de distinguer des sous espèces sont évidemment nettement plus grands chez les entomologistes.

Appliquées au cas qui nous occupe, les notions précédentes se traduisent de la manière suivante. Le flux de la parole est continu. Son analyse en termes de segments phonétiques discrets ne peut reposer ni sur des indices acoustiques, ni sur des indices articulatoires de discontinuité; elle ne peut reposer que sur une activité d'abstraction d'unités idéales. Une fois établies et au moins partiellement lexicalisées dans un alphabet, ces unités fournissent un moyen commode de description des énoncés de parole. Moyen commode, peut-être, mais aussi parfois source de malentendu tant il est difficile une fois de telles catégories établies de s'en détacher et d'adopter un point de vue descriptif différent. Il n'est donc pas étonnant que les informateurs de Sapir, connaissant une écriture alphabétique, puissent difficilement imaginer une orthographe basée sur un autre principe et témoignent d'une intuition plus

phonémique que phonétique. En effet, compte tenu du niveau d'expertise requis pour manipuler une écriture alphabétique, le phonème devient automatiquement le niveau fondamental de la catégorisation. En revanche, pour un phonéticien entraîné comme Sapir, disposant d'un alphabet plus précis et plus adapté à son niveau d'expertise -l'Alphabet Phonétique International- le niveau phonétique de description du langage devient sans doute le niveau fondamental de catégorisation.

Le point de vue qui vient d'être développé rejoint celui de Firth (1948). A l'instar des informateurs de Sapir qui ont été victimes de la tyrannie du segment phonémique résultant de leur connaissance de l'alphabet, la plupart des phonologues de ce siècle ont eux aussi été victimes de la tyrannie du segment phonétique à cause de leur décision d'adopter l'alphabet romain plutôt que l'alphabet grec comme base pour la transcription des sons élémentaires du langage. Firth déplorait qu'il en ait été ainsi et soulignait que la décision d'utiliser l'alphabet grec comme modèle aurait permis de représenter non seulement les aspects segmentaux mais aussi la prosodie grâce au système d'accents qu'il contient. Firth était convaincu que le segment phonétique pris comme entité n'est qu'une hypostase des l'alphabet romain.

Dans le même ordre d'idée, il faut souligner aussi que l'utilisation d'un outil descriptif fondé sur les segments peut être à l'origine d'erreurs dans l'établissement de la réalité psychologique des représentations inconscientes de constructs linguistiques (Mowrey & MacKay, 1990).

En effet, avec le développement des sciences cognitives depuis une quarantaine d'années, la question de la réalité psychologique des constructs linguistiques s'est déplacée; elle porte maintenant sur les représentations mentales inconscientes qui entrent en jeu dans le traitement de l'information linguistique. Ces représentations étant inaccessibles à l'introspection, leurs propriétés ne peuvent être qu'inférées. On dispose de deux sources principales d'inférence. La première source est constituée par les théories linguistiques, principalement les théories génératives développées par Chomsky et largement adoptées comme paradigme de la recherche linguistique depuis les années soixante. La seconde source d'inférence est empirique, fondée sur l'observation et l'expérimentation.

S'agissant de constructs phonologiques, comme les segments ou les traits distinctifs, l'analyse des erreurs d'élocution a constitué une des sources les plus importantes de démonstration de leur réalité psychologique en tant qu'unités intervenant dans la production et la perception du langage. Cette conclusion découle du fait que ces unités peuvent se transposer dans l'articulation d'un énoncé tout en préservant les contraintes phonotactiques de la langue. Or, ce que Mowrey et MacKay (1990) suggèrent de manière fort convaincante, c'est que des erreurs qui ne sont ni du niveau segmental, ni phonotactiques peuvent être idéalisées par l'observateur et forcées dans le moule offert par la transcription alphabétique dont il dispose. Ce qu'on a souvent interprété comme une démonstration de la réalité psychologique d'unités présidant à l'articulation du locuteur s'avère donc refléter la réalité psychologique de la normalisation des erreurs dans la description fournie par l'observateur.

2. DEFINITION ET CLASSIFICATION DES SYSTEMES D'ECRITURE.

Dans le présent exposé, la notion de système d'écriture s'applique exclusivement à des systèmes de notation glottographiques, c'est-à-dire des notations utilisant le medium visuel et permettant de transcrire de manière précise la succession des mots du discours. Sont donc exclus de cette définition des systèmes de notation éventuellement très précis, comme la notation mathématique ou le code de la route, dont les séquences de symboles peuvent être exprimées de plusieurs manières différentes (ex: *7 + 4* peut se lire "sept plus quatre", "sept et quatre" "l'addition de quatre et de sept", etc.). En simplifiant quelque peu, les diverses propositions de classification des systèmes d'écriture en fonction du type d'unité représentée par les signes graphiques peuvent assez bien se résumer dans une organisation tripartite hiérarchisée (Holender, 1987; Sampson, 1985). On distingue les systèmes logographiques qui représentent les morphèmes ou les mots des systèmes phonographiques qui représentent des unités phonologiques. Dans les systèmes phonographiques, il faut distinguer ceux qui reposent sur la syllabe de ceux qui reposent sur le phonème. Aucune orthographe n'est pure. Notamment, toutes les orthographes mélangent dans des proportions variables le principe logographique et phonographique. Ainsi, même dans des textes écrits de manière aussi alphabétique qu'en français, on peut trouver un certain nombre de logogrammes comme les chiffres arabes, des notations mathématiques (par exemple, %, #, +, -, x, :), des unités monétaires (par exemple, £, $) ou des signes typographiques (par exemple, &, §). Cette typologie traditionnelle des différents principes sur lesquels se fondent les écritures du monde n'est pas entièrement satisfaisante; elle appelle au moins les trois remarques suivantes.

Premièrement, on a longtemps eu tendance à envisager les systèmes d'écriture de manière évolutive. Gelb (1963) est sans doute un des derniers représentants de ce point de vue. Pour lui, l'évolution normale consiste à passer par les quatre stades suivants: idéographique, logographique, syllabique, alphabétique. Gelb pense que ne pas atteindre le stade alphabétique est un signe de blocage à un stade primitif et que l'atteinte d'un stade particulier ne peut s'effectuer que par un passage obligatoire par les stades antérieurs. Actuellement, plus personne ne pense comme cela. Le

stade idéographique ne fait pas à proprement parler de l'écriture définie comme une représentation fidèle de la chaîne du discours. L'écriture égyptienne est d'emblée logo-alphabétique et l'écriture sumérienne d'emblée logo-syllabique, du moins, comme on le verra plus loin, quand elle acquiert son caractère pleinement glottographique. Divers facteurs ont contribué au maintien de la conception évolutive de l'écriture. Le plus important est sans doute le fait que le caractère profondément linguistique de la transcription graphique du langage n'a pas été immédiatement entièrement perçu, notamment en raison de l'utilisation d'un vocabulaire partiellement inadapté pour décrire les signes graphiques. L'emploi de termes comme "pictogramme" et "idéogramme", par exemple, a contribué à occulter le caractère fondamentalement morphémique de l'unité de langage réellement représentée par les symboles ainsi désignés. En fait, l'élaboration d'un système d'écriture repose sur une analyse linguistique détaillée de la langue orale et il est maintenant bien établi que les différentes solutions adoptées sont en fait très bien adaptées aux diverses réalités linguistiques qu'elles représentent. Bien que le code alphabétique soit universellement applicable, il ne constitue nullement toujours la meilleure solution, ni le but ultime à atteindre.

Deuxièmement, tous les systèmes d'écriture représentent certains aspects de la substance phonique du langage. Ce fait très important est partiellement occulté par la distinction entre systèmes logographiques et systèmes phonographiques Prenons un exemple. Au moment de sa création, quand elle servait de moyen mnémotechnique pour enregistrer des transactions de type commercial, l'écriture sumérienne comprenait environ 1200 symboles dont la plupart sont des pictogrammes représentant des objets usuels (Powell, 1981). Pris isolément, il est difficile de décider du caractère plus ou moins idéographique ou logographique de tels symboles puisqu'ils représentent des concepts courants auxquels un nom est inévitablement associé. Le message transmis par l'arrangement de quelques uns de ces symboles, souvent combinés avec des nombres, peut avoir un sens très précis pour qui connaît les conventions du système tout en étant extrêmement indéterminé du point de vue de l'énoncé oral qui pourrait y être associé. Quelques siècles plus tard, quand le système a acquis sont caractère pleinement glottographique, permettant ainsi d'écrire de la poésie, par exemple, l'écriture cunéiforme est maintenant devenue logo-syllabique. Le répertoire des logogrammes s'est réduit à quelques centaines et un répertoire de signes syllabiques s'est développé à un tel point qu'il aurait pu être utilisé de manière autonome pour écrire le sumérien. Le maintient d'un répertoire de logogrammes est donc purement optionnel du point de vue de la transmission écrite des messages parlés. Tel est également le cas pour l'écriture logo-syllabique japonaise et pour l'écriture logo-alphabétique coréenne qui fonctionneraient tout aussi bien en ne conservant que leur partie phonographique.

Troisièmement, en corollaire du point 2, qu'en est-il alors de l'écriture chinoise, souvent considérée comme la seule à être quasi entièrement fondée sur le principe logographique? Ici aussi, le choix descriptif est source de confusion. La langue chinoise représentée dans l'écriture a pour particularité que tous les morphèmes sont monosyllabiques et invariables. Un caractère chinois représente une telle entité. La structure syllabique du chinois étant relativement simple, l'homophonie est fréquente. Une fois qu'un caractère a été inventé pour représenter un morphème, on peut le réutiliser pour représenter d'autres morphèmes homophones ou quasi-homophones. Pour éviter l'ambiguïté qui résulte immanquablement de ce processus, surtout quand les caractères sont isolés, on a eu recours à l'addition d'un complément sémantique, lui même puisé dans le répertoire existant. Le résultat est que la grande majorité des caractères chinois contiennent deux composantes, chaque composante étant elle même un caractère simple (non composé) du répertoire. Une des composantes indique un champ sémantique, par extension de la signification qui lui est associée quand elle fonctionne comme un caractère simple. L'autre composante doit être prise comme une indication globale de la prononciation du morphème monosyllabique transcrit par le caractère composé. Il est clair que la signification associée à cette composante quand elle correspond à une caractère simple ne joue aucun rôle fonctionnel dans la détermination du sens du caractère composé. Donc, chaque morphème est représenté par un caractère différent mais le nombre des symboles à apprendre est nettement moindre que le nombre de morphèmes. Le principe d'obtention de caractères composé par la combinaison d'une indication sémantique et d'une indication phonologique assure la productivité du système; on peut former des caractères nouveaux en combinant des caractères connus. On peut donc dire que le principe qui gouverne l'écriture chinoise est syllabique, même si, par opposition aux Sumériens, les Chinois n'ont pas développé de syllabaire, c'est-à-dire un répertoire de symboles systématiquement utilisés pour écrire les syllabes du chinois.

En résumé, les systèmes d'écriture sont des représentations linguistiques utilisant le medium visuel et permettant une transcription fidèle de la succession des mots du discours oral. Ces représentations sont fondées presque exclusivement sur la substance phonique -donc sur la composante phonologique- du langage. Seuls les phonèmes et les syllabes, ainsi que les mores en japonais, constituent les niveaux d'analyse phonologique représentés dans l'écriture (Holender, 1991; Mattingly, sous presse; DeFrancis, 1989).

3. ECRITURE IDEALE ET REALITE PSYCHOLOGIQUE DES UNITES DE TRAITEMENT.

Contrastons deux visions différentes de ce qui devrait idéalement être représenté dans l'écriture, celle de Gelb et celle de Chomsky. Gelb (1963, p. 246, ma traduction) pensait que:

"ce qu'il faut chercher est un système d'écriture combinant l'exactitude de l'alphabet API avec la simplicité formelle d'un système de sténographie".

Chomsky (1970, p. 12, ma traduction), en revanche, suggérait que:

"Une orthographe optimale, qui faciliterait l'utilisation des stratégies perceptives et des connaissances linguistiques disponibles, serait une orthographe qui entretiendrait une correspondance aussi étroite que possible, lettre-à-segment, avec la forme lexicale abstraite. Une telle orthographe conduit directement aux unités sémantiques et syntactiques significatives, s'abstrayant de toutes les propriétés phonétiques qui sont déterminées par des règles générales".

A la fois Gelb et Chomsky prônent donc une représentation écrite fondée sur un découpage du flux sonore continu en tranches verticales discrètes. Pour Gelb, les tranches verticales correspondent aux segments phonétiques et phonémiques de la phonologie structuraliste. Pour Chomsky, il s'agit des unités phonétiques et des segments phonologiques définis par Chomsky et Halle (1968) dans *The sound pattern of English* (SPE).

Ce que Gelb souhaite, c'est une représentation orthographique dans laquelle les lettres sont en correspondance biunivoque avec les phonèmes de la langue, un idéal qui est pratiquement réalisé en serbo-croate, par exemple. Les orthographes anglaise et française sont peu systématiques et irrégulières de ce point de vue. Notons que l'objectif visé par les orthographes alphabétiques au moment de leur conception a toujours été d'effectuer une transcription phonémique de la langue. Quand l'alphabet ne contient pas suffisamment de lettres pour représenter une langue non encore écrite, on a recours à des polygraphes ou à des signes diacritiques pour représenter les phonèmes manquants, ce qui est la preuve qu'on cherche à établir une correspondance entre les phonèmes et les graphèmes, le graphème étant défini comme une lettre ou un groupe de lettre représentant un phonème. C'est aussi quand l'orthographe s'écarte trop de cet idéal qu'on voit apparaître des gens qui préconisent des réformes tendant à restaurer la régularité des correspondances graphème-phonème.

Il faut souligner que la grande majorité des psychologues expérimentaux impliqués dans le débat sur les voies d'accès au lexique à partir de mots écrits alphabétiquement postule que la représentation phonologique lexicale est constituée par une séquence de phonèmes (Holender, 1988, pour une synthèse). Cette position impose une distinction entre des mots réguliers et irréguliers du point de vue de la correspondance entre graphèmes et phonèmes. Imaginons le processus d'accès au lexique pour ces deux types de mots. L'application de règles de correspondance graphème-phonème permet de synthétiser une représentation phonologique qui peut s'apparier à la représentation phonologique lexicale d'un mot régulier mais pas d'un mot irrégulier. Il faut donc imaginer que le lexique peut aussi être gagné de manière directe, sur base de l'information orthographique, puisque les mots irréguliers peuvent être lus. Cette procédure est évidemment également disponible pour les mots réguliers; ceux-ci peuvent donc accéder au lexique de deux manières différentes.

La position de Chomsky sur l'orthographe idéale découle de la manière dont la composante phonologique fonctionne en grammaire générative. Dans SPE, la seule unité prise en compte est le segment phonétique défini comme un vecteur de traits binaires, chaque trait pouvant être présent ou absent. Ces segments apparaissent dans la représentation phonologique et dans la représentation phonétique qui en est dérivée. Dans la représentation phonologique, chaque morphème est constitué par une chaîne de segments. Les morphèmes isolés sont stockés sous cette forme dans le lexique mental; ils apparaissent aussi sous cette forme dans la représentation de surface d'une phrase. Outre la succession des morphèmes, la représentation de surface comprend des indications correspondant aux frontières entre morphèmes et à la structure syntaxique de la phrase. Dans le processus de médiation entre le sens et le son, on peut dire que la représentation de surface est la sortie de la composante syntaxique et l'entrée de la composante phonologique. Les processus phonologiques consistent en une séquence ordonnée de règles qui opèrent sur la matrice des traits distinctifs de la représentation de surface, aboutissant à une représentation phonétique qui contient toute l'information nécessaire pour spécifier la prononciation de la phrase.

L'orthographe idéale, telle qu'elle est définie dans la citation de Chomsky mentionnée plus haut, est donc une orthographe qui établit une correspondance entre les graphèmes et les segments de la représentation lexicale plutôt que les segments de la représentation phonétique dérivée. En conséquence, l'orthographe de morphèmes prononcés différemment selon les mots dont ils font partie, comme par exemple HEAL et HEALTH, préservent l'invariance de la représentation morphémique sous-jacente en sacrifiant l'invariance de la correspondance graphème-phonème. De ce point de vue, l'orthographe anglaise est très proche de l'idéal.

Chomsky (1970) ajoutait que ce qui est vrai pour l'anglais en tant qu'orthographe quasi idéale l'est aussi pour les autres orthographes qu'il connaît. Or, ce point mérite d'être clarifié parce qu'un examen superficiel

d'orthographes telles que, par exemple, celle de l'italien ou du néerlandais et, a fortiori, celle du serbo-croate ou du finnois, semble démontrer à l'évidence que ce n'est pas la représentation lexicale abstraite mais la représentation phonétique qui sert de base à une transcription phonémique (au sens structuraliste). Les notions de profondeur phonologique et orthographique développées par Mattingly (1984) fournissent une solution à ce paradoxe.

Supposons que les langues diffèrent par la profondeur de leur phonologie, c'est-à-dire par le nombre de règles qui doivent être appliquées à la structure de surface pour engendrer la représentation phonétique. Pour une langue phonologiquement profonde, l'écart entre la représentation phonétique et la représentation lexicale sous-jacente sera grand parce qu'un grand nombre de règles auront été appliquées. En revanche, pour une langue phonologiquement superficielle, il y aura peu de différence entre la représentation sous-jacente et la représentation phonétique parce que la seconde aura été dérivée de la première par un faible nombre de règles. Par analogie avec la profondeur phonologique, on peut parler de la profondeur de l'orthographe. Une orthographe qui transcrit la représentation phonétique sera au degré minimum de profondeur alors qu'une orthographe qui transcrit la représentation lexicale sous-jacente sera d'autant plus profonde que la phonologie de la langue est plus profonde.

La thèse de Chomsky (1970), élaborée par Mattingly (1984), est qu'il n'existe pas d'orthographes superficielles parce que l'objectif est toujours de transcrire la représentation lexicale sous-jacente. Quand une orthographe paraît superficielle, c'est parce que la phonologie de la langue est superficielle. J'ai souligné ailleurs (Holender, 1987) que cette argumentation est restée complètement circulaire parce qu'aucune estimation de la profondeur phonologique des langues n'a été fournie pour appuyer l'hypothèse. En fait, il s'agit d'une question empirique. On ne peut rien conclure des langues à phonologie superficielle parce que même si elles tendent à réaliser l'idéal chomskien, les orthographes naturelles sont sans doute trop approximatives pour qu'on puisse discriminer entre la transcription de la représentation phonétique et de la représentation lexicale si celles-ci sont peu différentes l'une de l'autre. Par contre, en disposant d'un nombre suffisant de langues phonologiquement profondes et en inventoriant la fréquence avec laquelle leurs orthographes sont profondes aussi, on aurait la réponse à la question. Cette réponse ne viendra sans doute jamais parce que la question a pratiquement perdu son sens dans le cadre des théories phonologiques actuelles. Dans une analyse plus approfondie du problème (Holender, 1987), je spéculais sur le fait que la réponse aurait sans doute été opposée à l'hypothèse de Chomsky. La majorité des orthographes alphabétiques transcrivent la représentation phonétique à un niveau d'abstraction qui en font des représentations phonémiques se rapprochant bien plus de l'idéal de Gelb que de celui de Chomsky.

Chomsky a souvent décrété que les processus et les représentations mis en oeuvre par les différentes composantes du langage sont psychologiquement réels quoiqu'inconscients. Les connaissances qui entrent dans la composante phonologique, par exemple, sont des connaissances tacites quasi entièrement inaccessibles à l'analyse introspective mais dont la preuve de l'existence se manifeste dans la performance. Par exemple, tout locuteur peut juger si des logatomes constituent des candidats acceptables comme nouveaux mots de sa langue maternelle mais il ne peut généralement pas expliciter les règles phonotactiques qui justifient ses réponses. Dans la théorie phonologique développée dans SPE, le segment et le trait distinctif sont les seules unités qui soient considérées comme réelles du point de vue des processus psychologiques mis en jeu par la perception et la production du langage.

La position de Chomsky sur l'orthographe idéale conduit à un paradoxe. La représentation orthographique est une représentation consciente au sens où l'invention d'une représentation écrite et l'apprentissage de la lecture et des règles orthographiques sont des activités métalinguistiques. Or, Chomsky nous dit que la meilleure représentation orthographique possible doit être isomorphe avec la représentation lexicale sous-jacente qui est elle totalement inaccessible à la conscience. Le paradoxe est double. Le sujet conscient ne peut se rendre compte de l'isomorphisme entre les deux représentations puisque l'une d'entre elles est inaccessible à l'introspection. Le système de traitement ne le peut pas plus parce que les graphèmes de la représentation orthographique idéale ne sont que des symboles qui réfèrent à des vecteurs de traits, traits qui sont eux même abstraits du flux de la parole et spécifiés acoustiquement et/ou articulatoirement. Or, en admettant que les processus de traitement du langage oral résultent d'une évolution qui a permis que ces éléments primitifs soient extraits du flux de la parole, on voit mal comment ils pourraient l'être à partir d'une représentation orthographique dont l'isomorphisme est de second ordre. En effet, les symboles qui résument le contenu d'un vecteur de traits distinctifs n'ont de réalité que d'un point de vue métalinguistique conscient, il n'en ont aucune du point de vue des processus primaires de traitement du langage.

Mattingly (1972) a défendu une position qui avait le mérite d'éviter au moins partiellement le double paradoxe dont il vient d'être fait état. Il suggérait que la conscience linguistique trouve son origine dans un certain degré de pénétrabilité cognitive des processus primaires de traitement du langage. Cette propriété serait propre au langage et ne serait pas partagée par d'autres facultés cognitives, comme la perception de l'espace, par exemple. L'invention d'une écriture et l'apprentissage de la lecture et de l'orthographe dans un tel code sont des activités cognitives qui nécessitent une appréhension consciente des unités mises en jeu dans la représentation du langage. Or, si comme le suggérait

Mattingly à cette époque, cette conscience linguistique repose en partie sur l'analyse explicite des représentations sous-tendant le traitement du langage, il s'en suit que l'écriture alphabétique acquiert un statut particulier par rapport aux écritures fondées sur les syllabes ou les mores. Elle serait la seule à exploiter des unités qui sont psychologiquement réelles à la fois du point de vue des représentations mentales sous-jacentes au traitement du langage et du point de vue des représentations mentales conscientes. Toutefois, Mattingly (1984) a partiellement revu ses positions, suggérant que la conscience linguistique ne doit pas être envisagée comme la possibilité d'analyser introspectivement les représentations impliquées dans le traitement du langage mais simplement comme la possibilité d'y accéder. La preuve de cet accès se manifeste dans la performance, pas dans l'explicitation introspective de la connaissance qui est mise en oeuvre. Avec cette nouvelle formulation, dont on voit mal en quoi elle se distingue significativement de celle de Chomsky, on retombe dans le double paradoxe évoqué plus haut.

Pour conclure sur ce point, on peut dire que les théories phonologiques dominantes qui ont été développées depuis Saussure et jusqu'à il y a une quinzaine d'années sont des théories segmentales d'analyse de la substance phonique du langage. Le phonème dans sa définition structuraliste ou le segment phonétique tel qu'il est défini en grammaire générative, ainsi que le trait distinctif, sont les seules unités qui ont été jugées réelles du point de vue des processus de perception et de production de la parole. Des unités comme la syllabe, la more, les constituants syllabiques ou des aspects suprasegmentaux comme le ton sont indéniablement réels psychologiquement en tant que représentations mentales conscientes mais ils n'étaient pas jugés pertinents du point de vue du traitement du langage. Ce parti pris théorique a fortement influencé ce que les linguistes et, par la suite, les psycholinguistes, avaient à dire au sujet des systèmes d'écriture. Comme l'invention d'une orthographe repose à l'évidence sur une activité métalinguistique, on pourrait considérer que les différentes unités représentées dans les écritures ont toutes le même statut de descriptions conscientes alternatives du percept engendré par la parole. Toutefois, à cause du rôle théorique dominant joué par le segment, il était tentant de considérer que les écritures qui représentent ce niveau d'analyse, donc les écritures alphabétiques, ont un statut particulier en tant qu'elles jettent un pont entre les unités qui agissent comme des éléments primitifs du fonctionnement du langage et les unités orthographiques. C'est le point de vue qu'exprime si bien Studdert-Kennedy (1987, pp. 69, ma traduction) quand il dit que "historiquement, la possibilité de l'alphabet a été découverte, pas inventée".

4. DEVELOPPEMENTS RECENTS DES THEORIES PHONOLOGIQUES ET REALITE PSYCHOLOGIQUE DES UNITES ORTHOGRAPHIQUES.

Durant les quinze dernières années, les théories phonologiques se sont considérablement modifiées et diversifiées. Toutefois, l'existence de différentes options théoriques sous-jacentes à ces développements ne doit pas masquer les points de convergence qui sont nombreux. Parmi ceux-ci, j'en relèverai deux qui sont particulièrement importants pour le thème qui nous occupe: la multiplication des unités phonologiques jugées pertinentes pour le traitement du langage et la diminution du degré d'abstraction de ces unités.

Le premier point est peut-être le plus frappant. Le segment phonétique ou phonémique a perdu son statut privilégié qui en faisait pratiquement la seule unité phonologique pertinente théoriquement. Il n'est plus qu'un élément parmi d'autres, comme la syllabe, les constituants syllabiques, la more, ou le ton qui jouent maintenant des rôles fondamentaux dans l'analyse et la description de la composante phonologique du langage (voir, par exemple, Goldsmith, 1990).

Parmi les trois unités phonologiques représentées dans les systèmes d'écriture du monde -le phonème, la syllabe et la more- il n'en est plus une -le phonème- qui jouit d'un statut particulier par rapport aux deux autres. Toutes les trois sont maintenant considérées comme psychologiquement réelles à la fois en tant qu'unités descriptives disponibles à la conscience et en tant qu'unités de traitement impliquées dans la perception et la production du langage. Comme Studdert-Kennedy (1987) l'a bien souligné, le point de vue qu'il adopte à propos du phonème peut s'appliquer à d'autres unités utilisées pour représenter le langage dans l'écriture. A l'instar du phonème, la syllabe et la more sont aussi des unités implicites de la production de la parole qui attendent d'être découvertes, pas inventées, pour éventuellement être mises en rapport avec des symboles graphiques.

Ce point de vue est condamné à rester circulaire tant qu'on n'aura pas trouvé des arguments empiriques solides montrant que les unités de la description linguistique, psychologiquement réelles en tant que contenus de conscience accessibles à l'introspection et théoriquement pertinentes dans le système formel de description du fonctionnement du langage, sont aussi psychologiquement réelles en tant que représentations symboliques inconscientes présidant à la perception et à la production du langage. Or, nous avons vu que l'analyse des erreurs d'élocution, longtemps considérée comme fournissant la preuve empirique la plus convaincante de l'existence du segment et du trait phonétique en tant qu'unités de programmation de l'articulation, est illusoire. Mowrey et MacKay (1990) ont identifié la source de l'illusion; il s'agit de la tendance qu'a l'observateur à décrire les erreurs en termes des unités dont il cherche à démontrer l'existence.

Ceci nous ramène à nous interroger sur l'impact que les catégories discrètes fournies par le langage peut avoir sur notre conception du monde; en particulier sur les formulations théoriques se rapportant à des phénomènes fondamentalement continus et dynamiques. Cette question est d'autant plus difficile à aborder que, sous certains angles, les catégories discrètes du langage sont très éclairantes pour la compréhension de la nature des choses. Les catégories phonémiques, par exemple, sont particulièrement aptes à rendre compte des contrastes entre énoncés qui peuvent se produire d'un point de vue paradigmatique. En revanche, les segments phonétiques dénués de la dimension temporelle ne fournissent que des descriptions incomplètes, complexes et confuses sur le plan syntagmatique parce qu'ils représentent de manière statique des énoncés dont l'articulation et ses conséquences perceptives sont des phénomènes dynamiques.

Venons en alors au second aspect des théories phonologiques récentes qui nous intéresse, la diminution du degré d'abstraction des unités sous-jacentes. Sans entrer dans le détail, je pense qu'on peut affirmer avec Goldsmith (1990) que même dans les théories qui gardent leur caractère génératif, les unités phonologiques de la représentation lexicale, qui se sont maintenant multipliées, sont moins abstraites que dans SPE. Cette diminution générale du degré d'abstraction des représentations phonologiques est liée au fait que toutes les théories spécifient de manière de plus en plus adéquate les paramètres qui caractérisent l'articulation des énoncés. Mais la théorie qui y réussit le mieux est celle qui prend en compte les aspects dynamique de ce processus. Il s'agit de la théorie phonologique articulatoire développée par Browman et Goldstein (1986, 1991). Ici, la représentation lexicale consiste en une configuration de gestes articulatoires définis de manière dynamique. C'est à cause de différences dans le recouvrement entre gestes et éventuellement de différences dans le nombre de gestes impliqués que les items lexicaux diffèrent les uns des autres. On peut difficilement rêver d'une phonologie plus naturelle et plus concrète puisqu'elle se fonde immédiatement dans la production du langage et ses conséquences perceptives.

Dans la théorie articulatoire de Browman et Goldstein, les éléments primitifs de la représentation lexicale sont les gestes eux même. Il est évident que si l'on désire décrire le déroulement de gestes parallèles en saucissonnant la configuration gestuelle en tranches discrètes et en attribuant une valeur à chaque portion de geste ainsi obtenue, on va obtenir une matrice de traits distinctifs dont chaque vecteur vertical ressemble à un segment de la phonologie générative décrite dans SPE. En envisageant que les articulateurs responsables des différents gestes sont organisés de manière semi-hiérarchique et semi-autonome, on peut aussi construire des représentations qui ressembleront à celles de la phonologie autosegmentale. Tout ceci est parfaitement licite d'un point de vue descriptif. En revanche, si on souhaite modéliser la production du langage, donc la manière dont des configurations gestuelles différentes peuvent être engendrées, il est patent que ce qu'il faut comprendre c'est la manière dont les gestes articulatoires sont programmés et coordonnés. Il est peu probable qu'une caractérisation adéquate de ce processus puisse consister en des modifications locales des paramètres décrivant les portions de geste dans une matrice de traits distinctifs ou dans une représentation autosegmentale.

Le conclusion inévitable de ce point, qui sera aussi ma conclusion finale, est que si l'option théorique qui vient d'être développée est correcte, les unités phonologiques représentées dans les écritures et dans la plupart des théories phonologiques ne sont psychologiquement réelles que du point de vue des représentations mentales conscientes. Elles ne trouvent pas leurs correspondants sous forme d'unités plus ou moins isomorphes avec les unités conscientes dans les représentations mentales inconscientes qui sont utilisées par les processus primaires de production et de perception du langage. Contrairement à ce pense Studdert-Kennedy (1987). les unités phonologiques utilisées dans les différents systèmes d'écriture ont été inventées, pas découvertes.

REMERCIEMENTS.

Le présent travail a bénéficié de l'aide du Ministère Belge de la Politique Scientifique, Service de Programmation de la Politique Scientifique (Action de Recherche Concerté "Le traitement du langage dans différentes modalités: approches comparatives" Convention n° 91/96-148) et du Fonds de la Recherche Fondamentale Collective dans le cadre de la Convention n° 8.4505.92.

REFERENCES.

Browman, C. P., & Goldstein, L. M. (1986). Towards an articulatory phonology. *Phonology Yearbook, 3*, 219-252.

Browman, C. P., & Goldstein, L. M. (1991). Distinctiveness, phonological processes, and historical change. In I. G. Mattingly & M. Studdert-Kennedy (Eds.), *Modularity and the motor theory of speech perception* (pp. 313-336). Hillsdale, NJ: Lawrence Erlbaum .

Chomsky, N. (1970). Phonology and reading. In H. Levin & J. F. Williams (Eds.), *Basic studies in reading* (pp. 3-18). New York: Basic Books.

Chomsky, N. (1980). Rules and representations. *Behavioral and Brain Sciences, 3*, 1-61.

Chomsky, N., & Halle, M. (1968). *The sound pattern of English*. New York: Harper & Row.

DeFrancis, J. (1989). *Visible speech: The diverse oneness of writing systems*. Honolulu: University of Hawaï Press.

Firth, J. R. (1948). Sounds and Prosodies. *Transaction of the Philological Society*, 127-152. Reprinted In E. P. Hamp, W. Householder & R. Austerlitz (Eds.) (1966), *Reading in linguistics 2* (pp. 175-191). Chicago: University of Chicago Press.

Fodor, J. A. (1983). *The modularity of mind*. Cambridge, MA: The MIT Press.

Gelb, I. J. (1963, rev. ed.). *A study of writing*. Chicago: University of Chicago Press.

Goldsmith, J. A. (1990). *Autosegmental and metrical phonology*. Oxford: Basil Blackwell.

Holender, D. (1987). Synchronic description of present-day writing systems: Some implications for reading research. In J.K. O'Regan & A. Lévy-Schoen (Eds.), *Eye movements: From physiology to cognition* (pp. 397-420). Elsevier Science Publishers B.V. (North-Holland).

Holender, D.(1988). Représentations phonologiques dans la compréhension et dans la prononciation des mots. *Cahier du Département des Langues et des Sciences du Langage, 6., 31-84.*

Holender, D. (1991). Comment: Writing systems and the modularity of language. In I. G. Mattingly & M. Studdert-Kennedy (Eds.), *Modularity and the motor theory of speech perception.* (pp. 339-357). Hillsdale, NJ: Lawrence Erlbaum.

Hunt, E., & Agnoli, F. (1991). The Whorfian hypothesis: A cognitive psychology perspective. *Psychological Review, 98*, 377-389.

Liberman, A. M., & Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition, 21*, 1-36.

Liberman, A. M., & Mattingly, I. G. (1991). Modularity and the effects of experience. In R. R. Hoffman & D. S. Palermo (Eds.), *Cognition and the symbolic processes: Applied and ecological perspectives* (pp. 33-38). Hillsdale, NJ: Lawrence Erlbaum.

Mattingly, I. G. (1972). Reading, the linguistic process, and linguistic awareness. In J. F. Kavanagh & I. G. Mattingly (Eds.), *Language by ear and by eye* (pp. 133-147). Cambridge, MA: The MIT Press.

Mattingly, I. G. (1984). Reading, linguistic awareness, and language acquisition. In J. Downing & R. Valtin (Eds.), *Language awareness and learning to read* (pp. 9-25). New York: Springer Verlag.

Mattingly, I. G. (sous presse). Linguistic awareness and orthographic form. In R. Frost & L. Katz (Eds.). Hillsdale, NJ: Lawrence Erlbaum.

Mervis, C. B., & Rosch, E. (1981). Categorization of natural objects. *Annual Review of Psychology, 32*, 89-115.

Morais, J., Alegria, J., & Content, A. (1987). The relationships between segmental analysis and alphabetic literacy: An interactive view. *Cahiers de Psychologie Cognitive, 7*, 415-438.

Mowrey, R. A., & MacKay, I. R. A. (1990). Phonological primitives: Electromyographic speech error evidence. *Journal of the Acoustical Society of America, 88*, 1299-1312.

Powell, M. A. (1981). Three problems in the history of cuneiform writing: Origins, direction of script, literacy. *Visible Language, XV*, 419-440.

Rosch, E. (1973). Natural categories. *Cognitive Psychology, 4*, 328-350.

Sampson , G. (1985). *Writing systems: A linguistic introduction*. Stanford, CA: Stanford University Press.

Sapir, E. (1933). La réalité psychologique des phonèmes. *Journal de Psychologie Normale et Pathologique, 30*, 247-265.

Searle, J. R. (1990). Consciousness, explanatory inversion, and cognitive science. *Behavioral and Brain Sciences, 13*, 585-642.

Studdert-Kennedy, M. (1987). The phoneme as a perceptuomotor structure. In Allport, A., MacKay, D., Prinz, W., & Scheerer, E. (Eds.), *Language Perception and Production* (pp. 67-84). London: Academic Press.

Whorf, B. L. (1956). *Language, Thought, and Reality: Selected writings of Benjamin Lee Whorf* (edited by J. B. Carroll). New York: Wiley.

INTELLIGIBILITÉ COMPARÉE DU FRANÇAIS DE FRANCE À GRENOBLE ET À ABIDJAN

C. BENOIT [1], C. CHANARD [2], V. RISSOAN [1] & Z. TCHAGBALE [2]

[1] INSTITUT DE LA COMMUNICATION PARLÉE, U.A. CNRS N° 368, INPG/ENSERG - Université STENDHAL, BP 25X - 38040 GRENOBLE, FRANCE

[2] INSTITUT DE LINGUISTIQUE APPLIQUÉE, UNIVERSITÉ NATIONALE, 08 BP 887, ABIDJAN 08, République de CÔTE D'IVOIRE

Résumé

Le même test d'intelligibilité, subi à Grenoble par vingt auditeurs français, puis à Abidjan par vingt auditeurs ivoiriens, pour qui le français est une langue seconde, a permis de comparer les performances de deux groupes de sujets de même niveau d'étude mais aux compétences linguistiques différentes pour ce qui concerne la compréhension du français parlé en France, ou modélisé comme tel. Le test consistait, pour chaque sujet, en une transcription manuscrite de 500 phrases syntaxiquement simples, mais sémantiquement imprédictibles, et formées des monosyllabes les plus fréquents du français. Les stimulis avaient été synthétisés suivant deux méthodes de codage et deux modèles prosodiques élaborés pour le français "standard" d'une part, et prononcés par un locuteur français avant d'être spectralement dégradées d'autre part. Les scores d'intelligibilité globaux, en nombre de phrases correctement retranscrites, font apparaître une différence marquée entre les performances du groupe français et celles du groupe ivoirien. En revanche, la complexité linguistique perçue (mesurée comme le nombre d'unités de décision indépendantes nécessaires aux sujets pour la transcription de leur réponse : cf. l'article de Benoît aux précédentes JEPs) est indépendante de l'origine des sujets. Une telle observation est riche de conséquences pour les pédagogues comme pour les Industriels de la Langue Française ayant vocation à travailler avec ou pour la Francophonie en Afrique.

1. INTRODUCTION

Lors des précédentes JEPs (Benoît, 1990b), ont été présentés les résultats d'un test d'intelligibilité de la parole, synthétique et naturelle dégradée, réalisé auprès d'un échantillon de vingt auditeurs, tous étudiants français sans déficience auditive. Une mesure subjective de la complexité linguistique des phrases testées avait également été présentée L'indice proposé (rapport entre le logarithme des proportions de phrases correctes et celui des proportions de mots corrects) avait été utilisé antérieurement par Boothroyd (1988, 1990) pour quantifier une telle complexité, dans des phrases mais aussi dans des mots sans signification, tant auprès de sujets à audition normale qu'auprès de déficients auditifs. Ces travaux montrent que cet indice correspond au nombre d'unités de décision indépendantes traitées par les auditeurs pour tenter de comprendre une structure linguistique, mot ou phrase. En effet, la redondance introduite dans un message, compte tenu des contraintes au niveau du lexique, de la syntaxe, ou de la sémantique est une information mise à profit par les auditeurs pour réduire l'éventail des candidats possibles à la "compréhension" d'une unité linguistique (phonème, syllabe, ou mot, selon les cas et les niveaux de stratégie descendante). Ce phénomène se manifeste tout particulièrement quand la transmission acoustique du message est dégradée : la répartition des erreurs de transcription de messages suffisamment simples montre alors que les auditeurs traitent une structure de N unités linguistiques comme si elle n'était constituée que de P unités indépendantes (avec $P < N$), lorsque ces N unités sont reliées entre elles par un certain nombre de règles linguistiques (phonotactiques, lexicales, syntaxiques, etc.). La "redondance linguistique" du message peut donc être évaluée à $N - P$ unités indépendantes.

Dans cet article, nous comparons, entre deux communautés francophones géographiquement et culturellement disctinctes, non seulement l'intelligibilité globale de phrases transcrites manuellement, mais aussi la capacité des deux groupes d'auditeurs à tirer profit de cette "redondance linguistique" introduite, dans des phrases sémantiquement imprédictibles, par le respect de règles **lexicales** et **syntaxiques**.

Les phrases testées obéissent à cinq structures syntaxiques de base et ont été générées automatiquement par concaténation des monosyllabes français les plus

fréquents dans chacune des catégories grammaticales nécessaires. De telles phrases, appellées "sémantiquement imprédictibles" (Benoît et al., 1989 ; Grice, 1989 ; Benoît, 1990b), limitent l'aide à la compréhension des mots qui les constituent, lorsqu'elles sont présentées aux auditeurs sous forme acoustiquement dégradée, au seul respect de leur existence lexicale et de quelques règles syntaxiques de base en français. La dégradation que nous avons introduite dans ce test était soit due aux performances limitées de différents synthétiseurs de parole dans quatre cas, et à une dégradation artificielle du spectre de parole naturelle dans un cinquième cas. Les cinq conditions de "codage" utilisées permet de couvrir un champ suffisamment vaste de dégradation.

2. LES STIMULIS

Cinq structures syntaxiques ont été retenues, pour leur simplicité et pour leur adaptibilité, à quelques ajustements mineurs près, à l'ensemble des langues indo-européennes (Grice 1989). Des phrases ont ainsi pu être générées automatiquement à partir de lexiques grammaticaux contenant les monosyllabes les plus fréquents dans leur catégorie. Un exemple de "cadavre exquis" ainsi obtenu est présenté au tableau 1 ci-dessous pour chacune des cinq structures:

Tableau 1 : exemples de phrases utilisées dans le test

Structure 1 (forme intransitive)
La robe entre vers la science rouge.

Structure 2 (forme transitive)
Le verre vrai ouvre le coin.

Structure 3 (forme impérative)
Tourne la date ou la main.

Structure 4 (forme interrogative)
Quand le texte pose-t-il la fille crue ?

Structure 5 (forme relative)
La chose lance le train qui pense.

Vingt phrases ont été générées dans chacune des cinq structures. Les 100 phrases ont ensuite été synthétisées sous quatre conditions chacune : un synthétiseur à diphones codés à formants et un synthétiseur à diphones codés PSOLA, tous deux avec leur modélisation prosodique propre, et avec une prosodie constante. Elles ont aussi été lues par le locuteur dont la voix avait servi de référence pour la constitution des deux dictionnaires de diphones. Cinq bandes magnétiques ont alors été enregistrées. Chacune contenait les cinq groupes de vingt phrases d'une même structure syntaxique, présentées sous la même condition acoustique, après mise en ordre aléatoire de leur présentation. Un carré latin a permis d'aparier structures syntaxiques et conditions de codage de façon à ce que les cinq bandes magnétiques de 100 phrases chacune ne contiennent pas de répétition. Une pause de 10 secondes séparait deux phrases consécutives, et une pause d'une minute séparait deux groupes de vingt phrases.

3. LES AUDITEURS

A Grenoble comme à Abidjan, les auditeurs sélectionnés ne présentaient aucune déficience auditive connue. Ils étaient tous étudiants à l'université et avaient une bonne maîtrise de l'orthographe et de l'écriture manuscrite. Ils étaient rémunérés pour participer à l'expérience.

Le français était langue maternelle pour les 20 auditeurs grenoblois ; il était langue seconde pour les 20 auditeurs ivoiriens. Les auditeurs ivoiriens avaient tous poursuivi leur scolarité en français, langue véhiculaire officielle en Côte d'Ivoire.

Notons que, si le français est très répandu et parfaitement maîtrisé par les étudiants ivoiriens, sa spécificité régionale le rend de compréhension difficile pour un auditeur français peu habitué à cette variante. En revanche, les Ivoiriens sont beaucoup plus sensibilisés au français de France, du fait des médias en général, mais aussi de la présence française à l'Université pour les étudiants. Cette évidente dissymétrie de compréhension entre les deux variantes du français n'est pas étudiée ici, puisque nous avons seulement testé l'intelligibilité du français "de France" auprès des deux communautés.

4. LE TEST

Les deux tests se sont déroulés selon un protocole rigoureusement identique : cinq sous-groupes de quatre sujets ont subi cinq sessions au cours de chacune desquelles ils devaient retranscrire manuellement les 100 phrases entendues. Un carré latin avait réparti chacun des cinq groupes dans chacune des cinq sessions de façon à ce qu'une des cinq bandes magnétiques soient entendue à chaque session par un groupe, sans répétition. Les quatre auditeurs, équipés d'un casque de haute qualité, se tournaient le dos. Une brève session de familiarisation préalable leur permettait de prendre connaissance du genre de messages qu'ils avaient à retranscrire, puis du type de dégradation subie, sans autre précision sur le contenu, linguistique ou acoustique, du test. Il leur était recommandé d'être plutôt imaginatif en cas de mauvaise compréhension et de transcrire un trait quand un mot était jugé totalement incompréhensible.

5. LES RÉSULTATS

Les réponses aux 2 (tests) x 20 (auditeurs) x 5 (sessions, ou bandes magnétiques) x 100 phrases — soit 20 000 phrases, ou 130 000 mots environ — ont été saisies dans un fichier ASCII de façon à être analysées automatiquement en fonction de différents

facteurs. Nous renvoyons le lecteur à l'article antérieur de Benoît (1990b) pour une présentation détaillée des résultats chez les auditeurs français. Nous nous limitons ici à une comparaison des deux observations les plus significatives : l'intelligibilité globale des phrases et des mots pour chacun des auditeurs, moyennée sur l'ensemble des conditions acoustiques, syntaxiques et sur les sessions, d'une part ; et l'indice calculé comme le rapport des logarithmes des proportions précédentes, de façon à estimer la complexité linguistique perçue du corpus testé.

5.a Intelligibilité globale : comparaison des performances acoustiques

Nous avons mesuré l'intelligibilité globale moyenne des phrases composant l'ensemble du test, chez chacun des quarante auditeurs, comme le pourcentage d'unités correctement retranscrites. Comme unité, nous avons tout d'abord retenu le mot, celui-ci étant un monosyllabe dans chaque catégorie. Un mot a été considéré comme correct quand il était retranscrit à la bonne place dans la phrase, sous une forme orthographique telle qu'il pouvait appartenir à la catégorie grammaticale correspondante, avec les mêmes attributs morphologiques, et sous une forme homophonique. Ainsi, les mots *mer* ou *mère* étaient considérés comme intervertibles, alors que *mer* et *maire* ne l'étaient pas, ces derniers n'étant pas du même genre, bien qu'étant des noms tous les deux. Il faut noter ici que des mots susceptibles d'appartenir à deux catégories comme *vieux* (nom ou adjectif) n'avaient été retenus que dans le lexique de la catégorie la plus fréquente (adjectif, ici), de façon à désambiguïser l'appartenance d'un mot à la catégorie grammaticale prévue. Enfin, l'autre unité utilisée pour le critère d'intelligibilité était la phrase. Une phrase était considérée comme correcte si elle ne contenait aucun mot incorrect suivant le critère précédent.

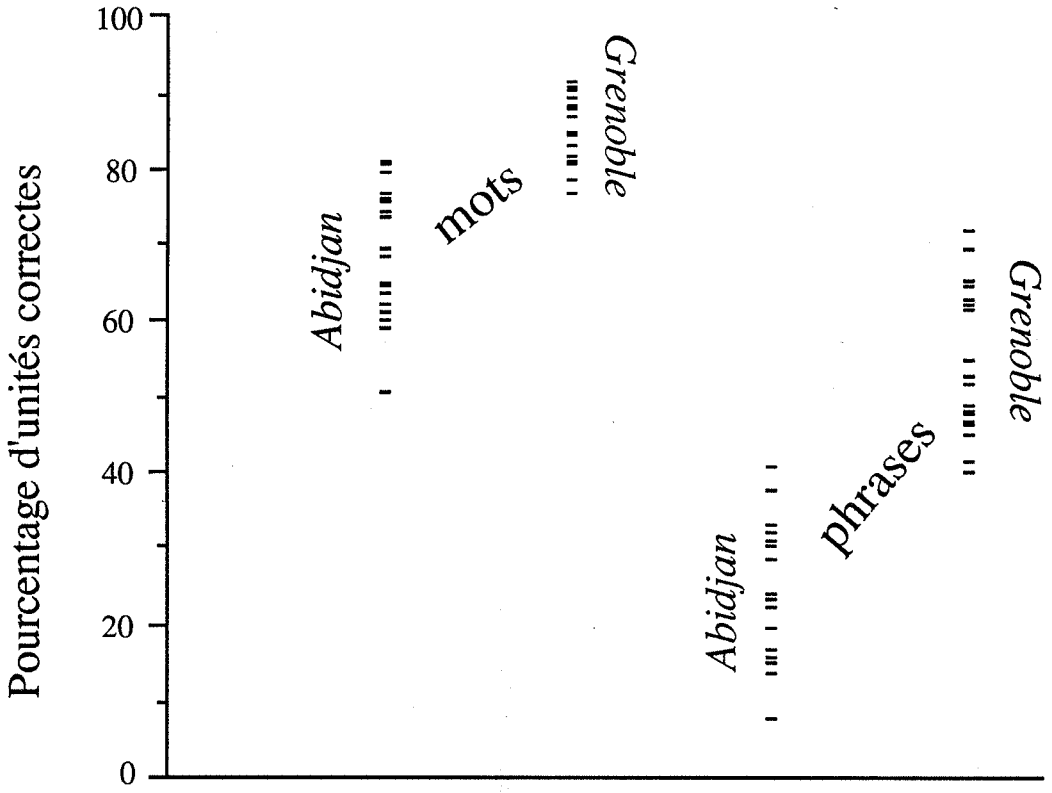


Figure 1. *Scores moyens d'intelligibilité, en pourcentage de mots corrects (à gauche) et de phrases correctes (à droite) obtenus par les vingt auditeurs ivoiriens (colonnes de gauche) et les vingt auditeurs français (colonnes de droite) sur les 500 phrases retranscrites.*

La Figure 1 présente les scores moyens d'intelligibilité en proportion de mots corrects et en proportion de phrases correctes chez les auditeurs ivoiriens et chez les auditeurs français. La différence de performance globale est très nette entre les deux groupes, puisque seul le "meilleur" des auditeurs ivoiriens obtient un pourcentage de phrases correctes supérieur à celui obtenu par le "pire" des auditeurs français.

Ce résultat global doit toutefois être nuancé en fonction de la dégradation acoustique, voire de la modélisation prosodique. Comme le montre la Figure 2,

la parole naturelle, assez faiblement dégradée, permet une discrimination totale entre les deux communautés, alors qu'un léger recouvrement est observé dans le cas du synthétiseur n° 1, et que le synthétiseur n° 2 confond encore plus nettement les deux communautés. Deux raisons peuvent être avancées a priori pour expliquer ce phénomène : les différences régionales d'intelligibilité sont atténuées soit quand l'acoustique est plus dégradée, soit quand la prosodie est moins naturelle. Or, en moyenne sur les 20 auditeurs ivoiriens, le pourcentage d'unités correctes n'est pas significativement modifié d'un modèle prosodique à l'autre (ou à son absence, plus exactement), quel que soit le synthétiseur. A l'inverse, chez les 20 auditeurs français, l'intelligibilité du synthétiseur n° 1 augmente de 58.0 % de phrases correctes, quand il est présenté en "prosodie constante", à 65.4 % quand il est présenté avec sa "prosodie modélisée", tandis que celle du synthétiseur n° 2 ne varie pratiquement pas d'une condition prosodique à l'autre. Comme la parole naturelle n'était présentée aux auditeurs qu'avec sa prosodie naturelle, robuste au type de dégradation acoustique infligé, il ne nous est pas permis de trancher dans cette énigme !

A travers les trois conditions acoustiques testées, il existe une étroite relation entre les qualités du codage acoustique et de la prosodie (du français parlé en France) associée : très bonnes en parole naturelle, médiocres avec le synthétiseur n° 1, et très pauvres avec le synthétiseur n° 2. C'est pourquoi nous nous limiterons ici à conclure, dans l'attente d'expériences complémentaires, soit qu'une forte dégradation acoustique ramène les deux groupes linguistiques à une incompréhension comparable des messages émis ; soit que c'est pour une très grande part l'information prosodique caractéristique d'une communauté de Francophones (ici celle de France) qui fait la différence d'intelligibilité entre deux messages, naturels ou synthétiques.

Il est légitime de supposer que les deux effets existent, et que ceux-ci se sont conjugués dans l'expérience rapportée ici.

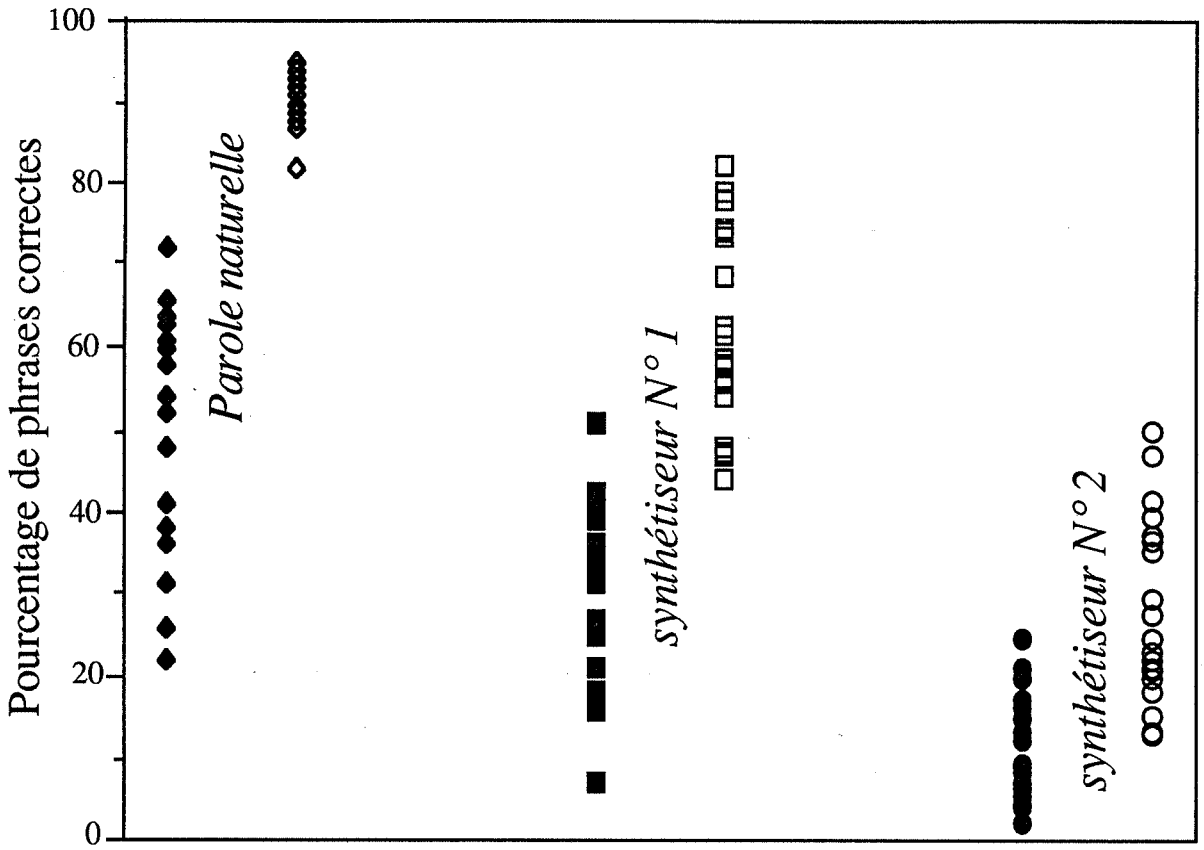


Figure 2. *Comparaison des pourcentages de phrases correctement retranscrites par les deux communautés selon le type de présentation acoustique: parole naturelle (faiblement) dégradée, et deux types de synthétiseur. Pour ces derniers, les résultats confondent les deux versions prosodie constante et prosodie modélisée.*

Un tel résultat doit évidemment être relativisé : les conditions expérimentales décrites ci-dessus sont (heureusement !) fort éloignées d'une situation de communication usuelle, dans laquelle les messages échangés entre interlocuteurs sont moins imprédictibles, chargés de sémantique et de pragmatique, d'une part, mais aussi moins acoustiquement dégradés ! Ainsi, dans des conditions aussi optimales que celles offertes par la télévision, il est permis de supposer que les messages diffusés "en français de France" aux auditeurs ivoiriens sont aussi bien compris que par les auditeurs français... En revanche, ce résultat établit clairement que les Industriels de la Langue et, avant eux, les chercheurs, doivent encore fournir un gros effort pour que leurs synthétiseurs de parole soient aussi compréhensibles par l'ensemble des utilisateurs francophones potentiels !

5.b Complexité perçue : comparaison des performances linguistiques

Une fois admises les différences intercommunautaires dans leurs performances de "bas niveaux", en matière de décodage acoustico-phonétique, et/ou en matière de décodage prosodique qui mettent en jeu des stratégies essentiellement ascendantes, il était intéressant de comparer leurs performances de plus "haut niveau", en matière de compétence linguistique mettant en jeu des stratégies descendantes.

Afin d'évaluer les différences cognitives des auditeurs, nous avons comparé la complexité linguistique des phrases perçue par chaque communauté. Pour ce faire, nous avons calculé, chez chaque auditeur, le rapport des logarithmes des deux proportions d'unités correctes r = Log (Pp) / Log (Pm).

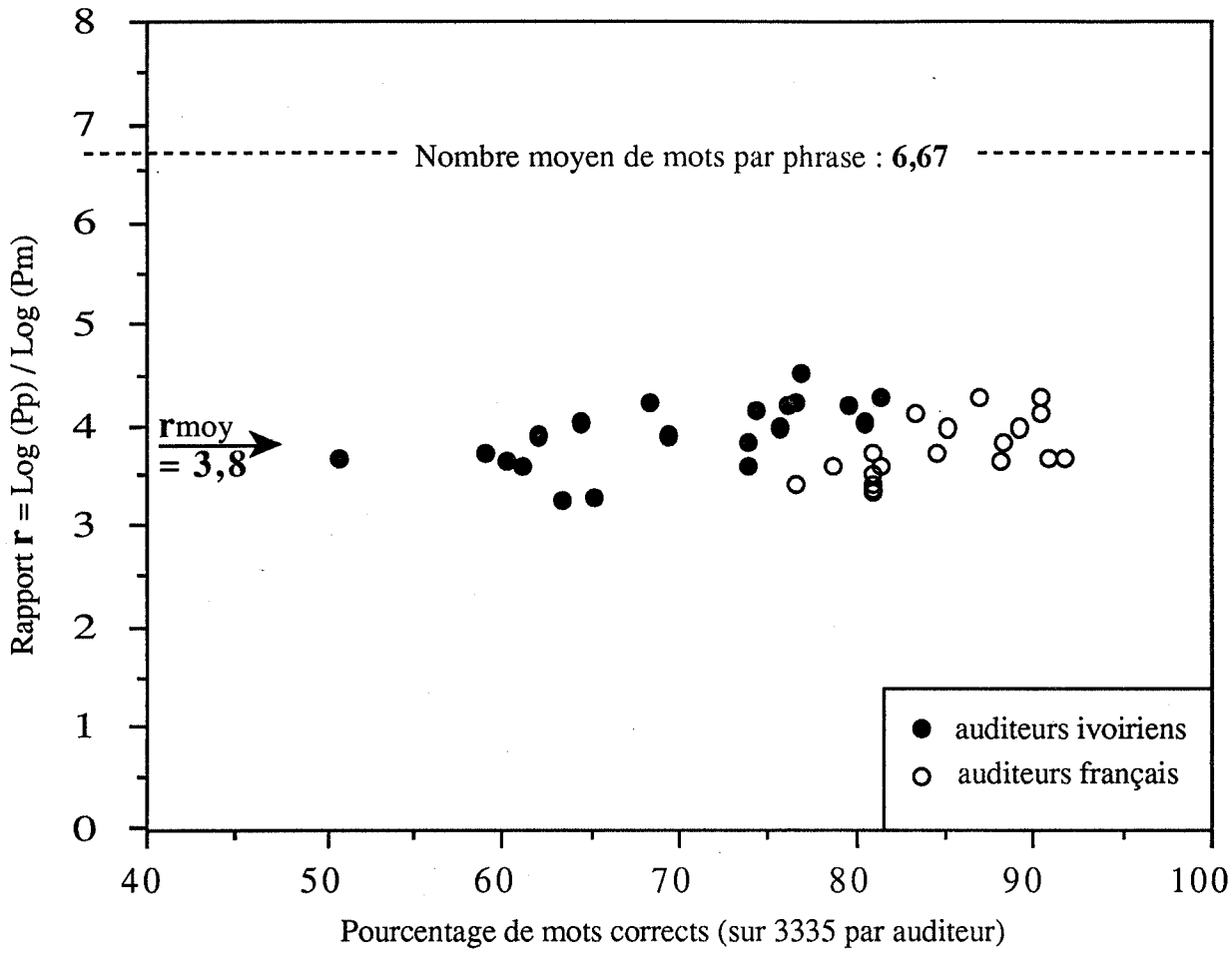


Figure 3. *Variation de "l'indice de complexité" r = Log (Pp) / Log (Pm) en fonction de l'intelligibilité globale en pourcentage de mots correctement retranscrits.*

Le rapport r est un indice robuste de la déviation entre i) la loi binomiale théorique des distributions d'erreurs sur des mots considérés comme indépendants les uns des autres et ii) la distribution observée des erreurs sur les mots des phrases testées, où des règles syntaxiques simples relient certaines unités entre elles : accord en genre entre le déterminant, l'adjectif et le nom ; accord morphologique entre ceux-ci et le verbe ; ordre d'apparition des mots dans la phrase ; etc.

Boothroyd (1988) et Benoît (1990b) ont montré que le rapport r = Log (Pp) / Log (Pm), dans lequel Pp représente la proportion de phrases correctes et Pm celle des mots corrects sur l'ensemble du test, représente le nombre d'unités de décision prises par un auditeur pour la compréhension des phrases d'un corpus homogène. Cet indice est sensible à de faibles variations des proportions. C'est pourquoi nous l'avons appliqué à l'ensemble des 500 phrases transcrites par chaque auditeur, sans distinction de structure syntaxique, afin d'en augmenter la précision. Il ne reflète donc pas la part prise par chacune des structures dans la "facilitation" à la compréhension des mots pris individuellement, mais l'avantage global apporté par l'ensemble des cinq structures à la compréhension de toutes les phrases. Cette approche holistique n'est pas choquante, dans la mesure où les cinq structures reflètent des syntaxes de base du français, et où les phrases ont sensiblement le même nombre de mots (6,67 en moyenne sur les 500 phrases).

La Figure 3 présente la valeur calculée de cet indice, pour chaque sujet, en fonction de la proportion de phrases correctement retranscrites.

La valeur du rapport des logarithmes reste à peu près constante d'un auditeur à l'autre (3,8 en moyenne contre 6,67 mots par phrase). Tous manifestent une large capacité à utiliser la redondance linguistique des phrases pour les traiter comme si elles n'étaient constituées que de 3,3 à 4,5 "unités indépendantes". Il n'apparaît donc pas de différence significative entre les deux communautés dans leur habileté à utiliser la redondance contextuelle des phrases, quelle que soit leur niveau d'intelligibilité.

7. CONCLUSION

Bien que les conditions dans lesquelles nous avons placé nos auditeurs étaient acoustiquement très dégradées et linguistiquement peu complexes, les observations tirées de l'expérience décrite dans cet article nous permettent d'affirmer que, **si les auditeurs ivoiriens ont plus difficulté que les auditeurs français à décoder acoustiquement – ou prosodiquement – des messages dégradés, les deux groupes font preuve d'une même compétence dans leur décodage linguistique.**

A notre connaissance, aucune étude comparative n'avait été menée jusqu'à présent sur l'intelligibilité comparée du français de France entre deux communautés francophones. Si de nombreux travaux ont été menés, dans les pays du Nord, sur l'analyse et la synthèse du français parlé au Québec, peu de cas est fait, dans la littérature, des variantes du Français parlé dans les pays du Sud. Or, il n'est pas utopique d'imaginer que l'utilisation de machines parlantes se répendra en Afrique aussi rapidement qu'en France ou au Canada, comme cela s'est déjà produit pour nombre de nouveautés technologiques.

N'ayons donc pas peur de saisir l'occasion des Journées d'Étude sur la Parole pour affirmer qu'il serait linguistiquement teinté d'un certain relent colonialiste francophone de la part des Industriels de la Langue – et plus encore, des chercheurs – si leurs synthétiseurs de parole considéraient davantage les variantes régionales du français comme l'objet d'un futur marché commercial que comme celui d'une étude linguistique. Ils développeraient alors plus un outil de normalisation linguistique qu'un support scientifique à l'enrichissment des connaissances universelles...

Remerciements

Cette étude a été partiellement subventionnée par le projet ESPRIT N° 2589, et par l'Agence de Coopération Culturelle et Technique.

Références bibliographiques

Benoît, C., **van Erp**, A., **Grice**, M., **Hazan**, V., & **Jekosh**, U. (1989), "Multilingual synthesizer assessment using semantically unpredictable sentences", *Proceedings of the Eurospeech '89 Conference*, Paris, 633-636.

Benoît, C. (1990a), "Mesure subjective de la redondance contextuelle : un indice pour quantifier la complexité linguistique", *Actes des 18e Journées d'Etude sur la Parole*, Montréal, Canada, 159-163.

Benoît, C. (1990b), "An intelligibility test using semantically unpredictable sentences: Towards the quantification of linguistic complexity", *Speech Communication*, 9, 293-304.

Boothroyd, A., & **Nittrouer**, S. (1988), "Mathematical treatment of context effects in phoneme and word recognition", *Journal of the Acoustical Society of America*, 84, 101-114.

Grice, M. (1989), "Syntactic structures and lexicon requirements for Semantically Unpredictable Sentences in a number of languages", Proceedings of the ESCA workshop on Speech I/O Assessment and Speech Databases, Noordwijkerhout, Hollande, 151-154.

Nittrouer, S., & **Boothroyd**, A. (1990), "Context effect in phoneme and word recognition by young children and older adults", *Journal of the Acoustical Society of America*, 87, 2705-2715.

19es J.E.P. - Bruxelles, 19 au 22 mai 1992.

UNE METHODE D'EVALUATION MULTICRITERE DE SORTIES VOCALES. APPLICATION AU TEST DE 4 SYSTEMES DE SYNTHESE A PARTIR DU TEXTE

M. CARTIER - F. EMERARD - D. PASCAL - P. COMBESCURE - A. SOUBIGOU

CENTRE NATIONAL D'ETUDES DES TELECOMMUNICATIONS
F-22300 LANNION

Résumé

On décrit une expérience d'évaluation de la qualité de parole de serveurs vocaux. La méthode utilisée est un essai d'écoute avec plusieurs échelles d'opinion : impression générale, effort d'écoute, problèmes de compréhension, articulation, prononciation, débit, agrément. Le corpus est constitué de 32 messages concernant la vente par correspondance ou la circulation des trains. Quatre systèmes de synthèse et trois références (une voix humaine et la même voix bruitée avec deux valeurs de signal-à-bruit) ont été testés conjointement. Les résultats montrent que cette méthode peut être fiable et discriminante.

INTRODUCTION

Peut-on, à partir d'essais en laboratoire, prévoir comment le client d'un serveur vocal appréciera la voix de la machine ? De nombreux outils d'évaluation existent, que se sont forgés les concepteurs de systèmes de synthèse, mais la plupart sont destinés au diagnostic et ne testent qu'un aspect particulier de la synthèse : "Evaluation can differ as a function of which TTS-system component one is interested in." (van Bezooijen et Pols, 1989).

Les problèmes de dialogue homme-machine et la variété des défauts inhérents à une réponse vocale automatique ne permettent pas d'appliquer à la parole synthétique les procédures adoptées pour la parole codée, mais l'évaluation de la qualité sonore d'une machine parlante présente quelques similitudes avec l'évaluation de la qualité de la parole comprimée. En effet, dans les deux cas : des sujets doivent écouter des signaux vocaux, puis transcrire des informations et/ou émettre des opinions ; les essais doivent fournir des données chiffrées qui permettront de situer le processus examiné par rapport à des critères donnés ou à des processus connus ; comme dans toute expérimentation subjective, les conditions physiques et psychologiques des tests doivent être suffisamment bien contrôlées pour que les résultats soient fiables et reproductibles.

Des études relatives à l'évaluation des codeurs ont montré qu'il était possible de réduire considérablement les aléas expérimentaux en testant conjointement des systèmes de référence et plusieurs systèmes à évaluer (Goodman & Nash, 1982). Des expériences d'évaluation de la parole synthétique, publiées dans des documents CCITT et résumées dans (Cartier, Karlsson et Modena, 1989), ont montré l'íntérêt d'une évaluation multicritère et la possibilité d'obtenir en laboratoire des résultats équivalents à ceux obtenus à l'aide d'essais en vraie grandeur.

A partir de ce double acquis et dans le but de disposer d'un outil de prévision de la qualité sonore des serveurs vocaux, une procédure d'évaluation multicritère portant sur plusieurs systèmes de synthèse et mettant en jeu des voix de référence a été élaborée. La présente communication la décrit et donne les résultats d'une expérience destinée à en vérifier la validité.

PRINCIPE

La procédure répond aux contraintes suivantes :

- elle peut s'appliquer à toute technique de production de parole ;
- elle place les sujets dans une situation sinon réelle, du moins proche d'une situation d'application ;
- plusieurs systèmes et références sont testés simultanément.

Les sujets entendent deux fois consécutives des messages susceptibles d'être émis par un serveur vocal. Lors de la première écoute ils transcrivent une partie des informations, lors de la seconde ils répondent à un questionnaire d'opinion comportant 4 questions. Les questionnaires sont de deux types :

Type I : Impression générale (G_i)
Effort d'écoute (I_1)
Difficultés de compréhension (I_2)
Articulation (I_3)

Type Q : Impression générale (G_q)
Qualité de prononciation (Q_1)
Débit (Q_2)
Agrément de la voix (Q_3)

I_1 porte sur l'effort nécessaire pour comprendre l'ensemble du message et en extraire les informations à transcrire, I_2 porte sur la compréhension des mots, I_3 sur la netteté de la prononciation. Q_1 est relative aux anomalies de prosodie, Q_2 à la vitesse d'élocution vis-à-vis de l'application, Q_3 à la voix telle qu'elle parvient au sujet.

SOURCES

Sept sources ont été testées, dont 4 systèmes de synthèse à partir du texte (à voix masculine) et 3 voix humaines : la voix d'un locuteur masculin et la même voix à laquelle a été superposé un bruit multiplicatif (simulant l'effet d'un bruit de quantification) avec 2 valeurs de rapport signal-à-bruit (RSB), 20 dB et 10 dB. Une valeur de 20 dB correspond à la limite inférieure de qualité, du point de vue du bruit de quantification, dans une communication téléphonique internationale. Toutes les sources ont émis la totalité du corpus ; elles ont été égalisées en niveau et filtrées à 3 400 Hz avant d'être présentées aux sujets à l'aide d'un combiné téléphonique (niveau d'écoute : 79 dB SPL). Dans ce qui suit, les systèmes de synthèse seront numérotés S_1, S_2, S_4 et S_6 ; la voix naturelle, S_5 ; les voix bruitées, S_7 (20 dB) et S_3 (10 dB).

CORPUS

L'ensemble du test concerne deux applications : vente par correspondance et circulation des trains. Voici un exemple de message pour chacune :

- *Madame Morin, la montre Data-Bank Casio 50 mémoires, bracelet résine, référence 811.19.04, au prix de 479 francs, vous sera livrée dans 3 semaines.*
- *Le train numéro 4119 en provenance d'Orléans arrivera à 12 heures 23, quai 8, voie H.*

Le corpus comporte 16 messages par application (14 pour le test proprement dit, 2 pour l'apprentissage). La durée moyenne d'un message est d'environ 20 secondes pour la vente et environ 12 secondes pour les trains.

Les messages de commande sont empruntés aux intitulés d'un catalogue de vente par correspondance, auxquels ont été ajoutés un délai de livraison et un nom patronymique tiré aléatoirement parmi les 55 noms français les plus fréquents. Les messages de circulation des trains sont proches de ceux diffusés dans les gares. Les 18 villes ont été choisies parmi les 50 villes françaises les plus importantes. Les messages ont été fabriqués automatiquement à partir d'un fichier texte ; des aménagements orthographiques ont été apportés pour obtenir une chaîne phonétique correcte.

PLAN D'EXPERIENCE

Un certain nombre de contraintes ont été fixées pour l'ensemble du test :

- L'objectif étant de prédire la qualité d'un serveur vocal dans le contexte d'une application déterminée, la totalité des messages relatifs à une application doit être présentée à l'intérieur d'une seule séance ou d'un seul sous-ensemble de séance.
- Afin d'éviter tout artéfact de corrélation entre les deux séries de jugements, la collecte des réponses aux questionnaires I et aux questionnaires Q doit être séparée.
- Le plan d'expérience doit permettre de contrôler un éventuel effet message, car on est loin d'être assuré de l'intelligibilité parfaite de la parole synthétique.

Nous avons opté pour les solutions suivantes :

- L'expérience a été découpée en deux sessions, une par application, séparées par une pause. Tous les sujets ont entendu les messages de vente par correspondance (C) pendant la première session et les messages de trains (G) pendant la seconde. Ils étaient ainsi parfaitement "naïfs" pour l'application C , et la légère familiarisation acquise au bénéfice de l'application G était la même pour tous.

- Chaque session est constituée de deux blocs. Dans le premier bloc, les sujets ont eu à remplir des questionnaires I ($G_i + I_1 + I_2 + I_3$) ; dans le second , des questionnaires Q ($G_q + Q_1 + Q_2 + Q_3$). L'ordre de collecte des évaluations a été inversé pour les deux applications : 7 questionnaires I puis 7 questionnaires Q pour la première, 7 questionnaires Q puis 7 questionnaires I pour la seconde.

- Les paramètres variables étant au nombre de quatre (sources, messages, ordres de présentation, sujets), nous avons construit l'expérience à partir d'un plan de base en carré gréco-latin. Ce plan permet de tenir compte de 4 facteurs de variabilité, le facteur principal à 7 modalités (4 systèmes de synthèse + 3 voix naturelles) et 3 facteurs concomitants (messages, ordres, sujets) de modalités égales elles aussi à 7.

Ce type de plan permet de réduire à 49 (7^2) le nombre d'essais nécessaires à l'estimation des 4 effets principaux. Le présupposé à la base de ce schéma expérimental est qu'il n'existe pas d'interactions entre les 4 facteurs (sources, sujets, messages et ordres). En particulier, avec un tel plan il n'est pas possible d'estimer l'effet de l'interaction source-message. Il faut donc s'assurer au préalable, par un bon échantillonnage de la population des sujets et un choix de messages homogènes, que ces interactions sont, par construction, négligeables.

Au total quatre blocs indépendants ont été constitués :

bloc 1 : application C, questionnaires I
bloc 2 : application C, questionnaires Q
bloc 3 : application G, questionnaires Q
bloc 4 : application G, questionnaires I.

Chaque bloc a été construit à partir de carrés gréco-latins 7 x 7 différents. L'organisation de la première session est indiquée ci-dessous ; les colonnes représentent les groupes sujets et les lignes, les ordres de présentation. S_i identifie les sources (i = 1 à 7) et C_j les messages vocaux (j = 1 à 14) :

Grp1	Grp2	Grp3	Grp4	Grp5	Grp6	Grp7
S C	S C	S C	S C	S C	S C	S C
2 2	3 3	4 4	5 5	6 6	7 7	1 1
3 6	4 7	5 1	6 2	7 3	1 4	2 5
5 7	6 1	7 2	1 3	2 4	3 5	4 6
4 3	5 4	6 5	7 6	1 7	2 1	3 2
7 1	1 2	2 3	3 4	4 5	5 6	6 7
1 5	2 6	3 7	4 1	5 2	6 3	7 4
6 4	7 5	1 6	2 7	3 1	4 2	5 3
3 10	4 11	5 12	6 13	7 14	1 8	2 9
4 14	5 8	6 9	7 10	1 11	2 12	3 13
6 8	7 9	1 10	2 11	3 12	4 13	5 14
5 11	6 12	7 13	1 14	2 8	3 9	4 10
1 9	2 10	3 11	4 12	5 13	6 14	7 8
2 13	3 14	4 8	5 9	6 10	7 11	1 12
7 12	1 13	2 14	3 8	4 9	5 10	6 11

Plan d'expérience, blocs 1 et 2
(7 questionnaires I puis 7 questionnaires Q)

Les messages utilisés pour le questionnaire I (C1 à C7, G8 à G14) étaient différents de ceux utilisés pour le questionnaire Q (C8 à C14, G1 à G7). Pour disposer d'une meilleure base statistique, le nombre de sujets a été multiplié par 4, chaque colonne du carré représentant un groupe de 4 sujets. Au total 28 sujets (7 groupes de 4) ont participé à l'expérience (étudiants, rémunérés, 13 femmes et 15 hommes).

L'ensemble du test s'est déroulé en 7 séances d'une heure, instructions comprises. Avant l'écoute des 28 messages du test proprement dit, on a procédé à la lecture à haute voix de consignes écrites et les sujets ont rempli 6 feuilles de réponse portant sur des messages d'apprentissage qui leur fournissaient un éventail des conditions et leur donnaient l'occasion de répondre aux différentes questions.

Transcriptions

Chaque message donnait lieu à la transcription de cinq réponses. Pour les messages C, c'étaient le nom du client, la désignation de l'article (limitée à 3 mots maximum), la référence (3 nombres inférieurs respectivement à mille, cent et cent), le prix et le délai (un nombre chacun) ; pour les messages G, le numéro du train, le nom de la ville, l'horaire, le quai et la voie. On a comptabilisé une faute pour toute réponse incomplète ou erronée quels que soient la nature et le nombre d'erreurs dans la réponse.

Echelles d'opinion

Toutes les échelles comportent cinq possibilités de réponse (par exemple, les réponses de l'échelle de difficulté de compréhension -"*Avez-vous éprouvé de la difficulté à comprendre certains mots ?*"- sont : jamais, rarement, de temps en temps, souvent, tout le temps). Ces réponses ont été codées de 5 à 1 en allant de la meilleure à la moins bonne, sauf celles de l'échelle de débit ("*Le débit vous convient-il ?*" : oui ; oui mais un peu lent ; oui mais un peu rapide ; non, trop lent ; non, trop rapide), dont ont été dérivées une note de "qualité" Q2q (resp. 5 ; 3,5 ; 3,5 ; 2 ; 2) et une note de "vitesse" Q2v (5, 4, 6, 3, 7).

RESULTATS

Le tableau ci-dessous donne le nombre de fautes par source et par application. Chaque nombre de fautes correspond à 280 réponses (5 réponses x 2 messages x 28 auditeurs) :

	S1	S2	S3	S4	S5	S6	S7
C	17	43	20	58	2	67	7
G	14	21	19	19	4	14	10

Fautes de transcription

Le tableau 1 donne les notes moyennes d'opinion (MOS, mean opinion scores). La figure 1 présente une partie de ces MOS sous forme de graphiques.

On observe que la voix naturelle fortement bruitée (S_3) est notée plus favorablement en débit que la même voix non bruitée (S_5).

Les analyses statistiques effectuées (bloc par bloc) portent sur les notes d'opinion. Le tableau 2 donne le résultat des analyses de variance effectuées pour chaque note d'opinion de l'application C. Le tableau 3 donne le résultat de tests HSD (honestly significant differences ; Edwards, 1972) pour une partie de ces données.

On constate que l'effet auditeur est peu important mais significatif. L'effet ordre et l'effet message sont négligeables.

L'effet source est prépondérant. Les tests HSD montrent que cet effet correspond effectivement à une bonne discrimination entre sources. On notera en particulier les différences significatives entre la voix naturelle (S_5) et la voix bruitée à 20 dB (S_7).

CONCLUSION

A l'examen des résultats, on peut conclure que :

- la procédure est fiable et discriminante,
- l'utilisation de plusieurs échelles paraît justifiée.

La reproductibilité interlaboratoire reste à démontrer et, bien entendu, des essais en vraie grandeur seront nécessaires pour transformer cet outil de laboratoire en un instrument de prévision de qualité et d'acceptabilité.

BIBLIOGRAPHIE

van Bezooijen R. & Pols L.C.W. (1990), "Evaluating text-to-speech systems", *Speech Communication,* vol. 9, n° 4, 263-270.

Cartier M., Karlsson C. & Modena G. (1989), "Standardization of synthetic speech quality assessment for telecommunication purposes", *Proceedings ESCA Workshop Speech I/O Assessment and Speech Databases, Noordwijkerhout,* 3.5.1-3.5.4.

Edwards A.L. (1972), *Experimental design in psychological research,* éd. Holt, Reinhart & Winston, Inc.

Goodman D.J. & Nash R. (1982), "Subjective quality of the same transmission conditions in seven different countries", *Proceedings ICASSP, Paris,* 984-987.

Echelle	Bloc		S1	S2	S3	S4	S5	S6	S7
Impr. générale	1	Gi	3,32	2,46	2,18	1,96	4,71	1,96	3,32
Impr. générale	2	Gq	3,29	2,21	2,21	1,96	4,64	1,68	3,54
Effort d'écoute	1	I1	3,93	3,25	3,61	2,96	4,96	2,61	4,43
Compréhension	1	I2	3,86	2,96	3,68	2,57	4,93	2,54	4,54
Articulation	1	I3	3,29	2,75	3,71	2,18	5,00	2,07	4,32
Prononciation	2	Q1	3,57	2,96	4,04	2,43	4,96	1,96	4,54
Débit ("qualité")	2	Q2q	4,04	2,80	4,68	3,98	4,14	2,91	4,46
Agrément voix	2	Q3	2,93	1,71	2,79	1,89	4,46	1,61	3,71
Débit ("vitesse")	2	Q2v	5,43	6,25	4,79	5,46	4,57	6,18	4,79

C
(commandes)

Echelle	Bloc		S1	S2	S3	S4	S5	S6	S7
Impr. générale	4	Gi	3,50	2,54	1,79	2,54	4,64	2,46	3,39
Impr. générale	3	Gq	3,32	2,43	2,25	2,75	4,71	2,36	3,68
Effort d'écoute	4	I1	3,93	3,29	3,43	3,29	4,89	3,04	4,64
Compréhension	4	I2	4,04	3,46	3,68	3,21	4,93	3,11	4,57
Articulation	4	I3	3,43	2,71	3,39	2,32	4,82	2,50	4,11
Prononciation	3	Q1	3,68	2,82	3,89	3,32	4,96	2,68	4,75
Débit ("qualité")	3	Q2q	3,71	2,64	4,84	3,71	4,52	3,02	4,52
Agrément voix	3	Q3	2,96	1,93	3,00	2,39	4,54	2,14	3,82
Débit ("vitesse")	3	Q2v	5,86	6,57	4,96	5,79	4,82	6,32	4,68

G
(trains)

Tableau 1 - MOS (notes moyennes d'opinion)

			Gi imp. gén.	Gq imp. gén.	I1 effort	I2 compr.	I3 artic.	Q1 pron.	Q2q débit	Q3 agr.	Q2v débit
Total	195 df	MSq	1,50	1,60	0,91	1,25	1,53	1,77	1,24	1,48	1,01
Residu	150 df	MSqrés	0,50	0,50	0,24	0,37	0,43	0,58	0,75	0,34	0,44
Sujets	Aud/Rés	F1	2,62 **	2,97 **	3,37 **	3,36 **	2,50 **	3,06 **	1,34	3,98 **	3,05 **
Ordres	Ord/Rés	F2	2,40 *	1,33	1,44	1,52	0,27	0,90	0,78	0,80	2,57 *
Conditions	Syn/Rés	F3	56,35 **	64,05 **	80,05 **	66,08 **	77,29 **	59,16 **	19,86 **	95,92 **	29,14 **
Messages	Mes/Rés	F4	1,43	1,11	0,84	0,69	0,76	1,00	2,43 *	1,39	3,76 **

Tableau 2 - Analyse de la variance, messages C (commandes, blocs 1 and 2)

** : significatif à .01
* : significatif à .05

Gi	S5	S1	S7	S2	S3	S4	S6
S5		**	**	**	**	**	**
S1	**		=	**	**	**	**
S7	**	=		**	**	**	**
S2	**	**	**		=	=	=
S3	**	**	**	=		=	=
S4	**	**	**	=	=		
S6	**	**	**	=	=	=	

Q3	S5	S7	S1	S3	S4	S2	S6
S5		**	**	**	**	**	**
S7	**		**	**	**	**	**
S1	**	**		=	**	**	**
S3	**	**	=		**	**	**
S4	**	**	**	**		=	=
S2	**	**	**	**	=		=
S6	**	**	**	**	=	=	

** : significatif à .01
= : non significatif à .05

I1	S5	S7	S1	S3	S2	S4	S6
S5		**	**	**	**	**	**
S7	**		**	**	**	**	**
S1	**	**		=	**	**	**
S3	**	**	=		=	**	**
S2	**	**	**	=		=	**
S4	**	**	**	**	=		=
S6	**	**	**	**	**	=	

I2	S5	S7	S1	S3	S2	S4	S6
S5		=	**	**	**	**	**
S7	=		**	**	**	**	**
S1	**	**		=	**	**	**
S3	**	**	=		**	**	**
S2	**	**	**	**		=	=
S4	**	**	**	**	=		=
S6	**	**	**	**	=	=	

Tableau 3 - Test HSD pour 4 MOS, messages C (commandes, blocs 1 et 2)

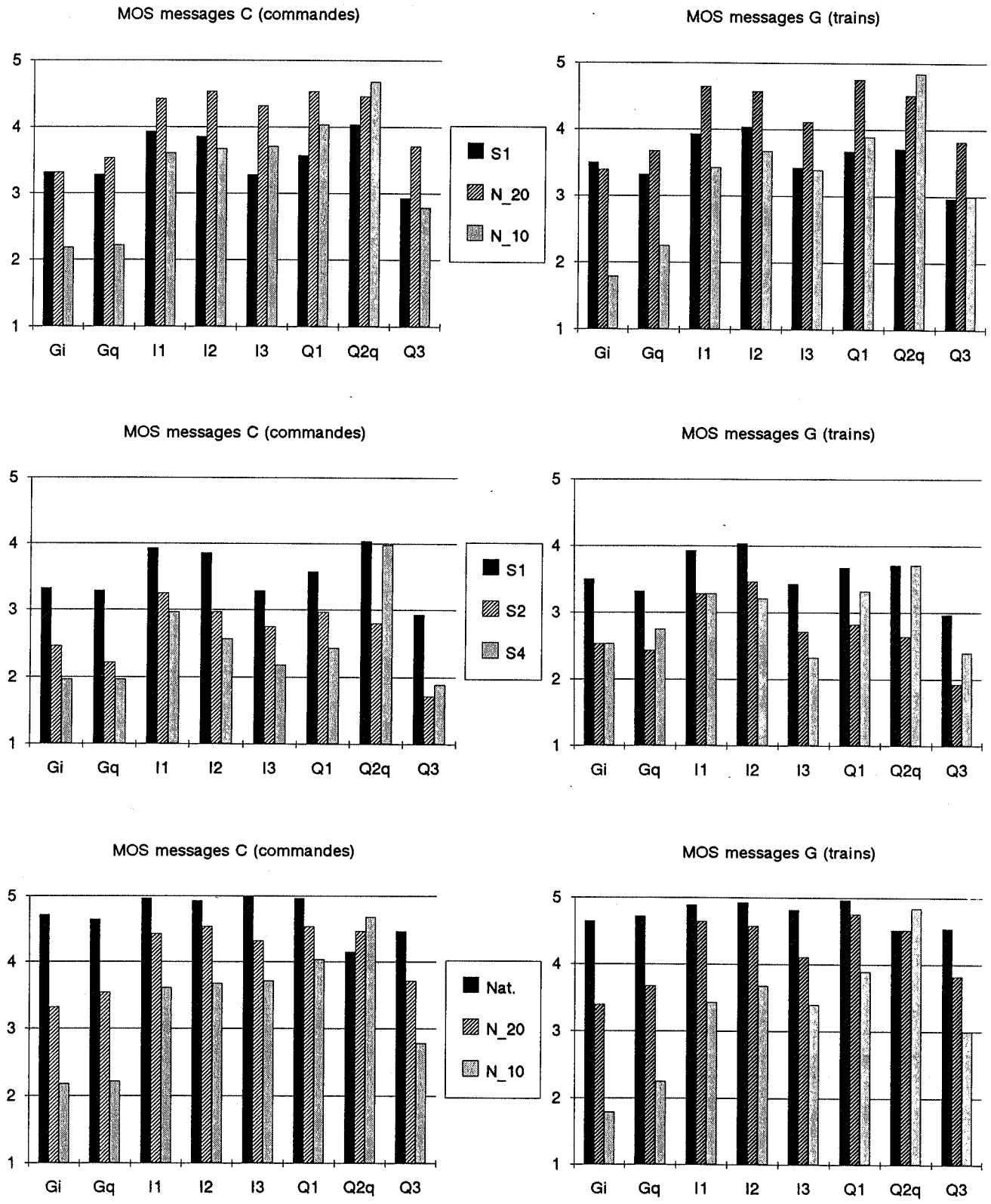


Fig. 1 - Notes moyennes d'opinion comparées : de haut en bas, 1 synth. et 2 nat., 3 synth., 3 nat.

19es J.E.P. - Bruxelles, 19 au 22 mai 1992.

PERCEPTION DES TRAITS PHONETIQUES: EFFET DU CONTEXTE SUR L'INTEGRATION DES INDICES ACOUSTIQUES

WILLY SERNICLAES

INSTITUT DE PHONETIQUE ET ECOLE DE SANTE PUBLIQUE
UNIVERSITE LIBRE DE BRUXELLES

Résumé

Des travaux antérieurs suggèrent que le contexte vocalique a des effets additifs sur la perception des fricatives en anglais. Mais les effets de la voyelle sur la production des fricatives sont également additifs dans ce cas, et l'absence d'interaction perceptive devient dès lors triviale. L'expérience présentée ici concerne les relations entre fricatives et voyelles en français et fournit un contre-exemple en suggérant qu'une interaction dans la production des indices et des traits s'accompagne effectivement d'une interaction perceptive.

1. INTRODUCTION

On sait que la perception d'un trait phonétique peut, en cas d'ambiguïté acoustique, dépendre de l'identité des traits contextuels (Refs.1,2). L'effet du contexte sur la perception d'un trait pourrait soit provenir d'un biais dans l'évaluation des indices acoustiques correspondants, soit d'un ajustement des règles d'évaluation de ces indices. Des travaux antérieurs suggèrent que l'on se trouve en présence de biais (Refs.3,4). Les indices manipulés dans ces travaux ne fournissent cependant pas d'information pertinente pour opérer un traitement interactif (Ref.5).

Prennons l'exemple de l'interdépendance entre l'identification du lieu d'articulation des fricatives et de la voyelle pour des syllabes CV en anglais. Avec des stimuli ambigus entre [si, su, chi, chu], les sujets fournissent davantage de réponses [s] lorsque la voyelle est perçue comme [u], et inversément (Ref.6). Il s'agit d'un effet additif car il ne varie pas en fonction des fréquences des formants (Ref.3). Mais ceci n'est pas étonnant si l'on sait que les différences entre consonnes [s-ch] et entre voyelles [i-u] ont également des effets additifs sur la *production* des formants (Ref.5).

L'information fournie par ces indices sur l'une des deux oppositions ne dépend donc pas de l'autre opposition et elle n'est donc pas susceptible d'un traitement interactif, du moins dans une optique fonctionnelle (Ref.8).

Le problème qui se pose est de tester l'hypothèse

d'ajustement à l'aide d'indices et de traits adéquats. Le travail présenté ici, et qui concerne les relations entre les fricatives [f, s , ch] et les voyelles [i, y, u] du français, constitue un premier pas dans cette direction.

2. MESURES ACOUSTIQUES

2.1 Procédure

Un ensemble de 18 syllabes (fricative [f,s,ch,v,z,j]+ voyelle [i,y,u]) a été prononcé par 4 sujets francophones masculins. Le locuteur était placé en chambre anéchoïque et les syllabes étaient enregistrées sur bande analogique.
Les fréquences des F2 et F3 ont été mesurés sur spectrogrammes analogiques (wide 300 Hz) en 3 moments dans le déroulement temporel de la syllabe. Comme les formants n'étaient pas toujours visibles, la localisation des points de mesure est approximative. La premiére mesure spectrale a été prélevée aussi près que possible du début du bruit de friction, la deuxième aussi prés que possible de la fin du bruit et la troisième à la fin des transitions vocaliques.
Sur l'ensemble des 432 mesures projetées (18*4*2*3) seules 316 (73%) ont pu être effectuées, en raison du manque de visibilité des formants.

2.2 Résultats acoustiques

Les fréquences moyennes des F2 et F3, regroupées sur les 4 locuteurs et les 2 catégories de voisement, sont présentées dans le Tab.1. Pour chaque syllabe et chaque formant, 2 mesures sont fournies, l'une correspondant à la moyenne des mesures effectuées au début et à la fin du bruit de friction et l'autre aux mesures effectuées à la fin des transitions vocaliques.
Les valeurs des F2 et F3 relatives à la friction, d'une part, et au segment vocalique, d'autre part, ont été traitées séparément par un plan d'Anova à 3 facteurs (consonne, voyelle, formant). Pour la friction, les effets significatifs intéressants (pour notre propos) sont ceux de la consonne ($p=.013$), de la voyelle ($p<.001$), et surtout celui de l'interaction voyelle-consonne ($p=.031$).
Pour le segment vocalique, les effets de la consonne ($p=.006$) et de la voyelle ($p<.001$) sont également significatifs, tandis que leur interaction ne l'est pas ($p=.587$).

Tab.1 Mesures acoustiques (Hz) des F2 et F3 durant la friction (fri) et le segment vocalique (voc).

	[i]	[y]	[u]
[f+ v]			
F2 fri	1669	1577	1264
voc	1866	1613	711
[f+v]			
F3 fri	2394	2251	2208
voc	2763	2090	2066
[s+ z]			
F2 fri	1677	1652	1314
voc	1909	1674	802
[s+ z]			
F3 fri	2511	2424	2297
voc	2828	2188	2084
[ch+j]			
F2 fri	1780	1554	1309
voc	1926	1613	808
[ch+j]			
F3 fri	2469	2191	2380
voc	2816	2081	2114

3. EXPERIENCE DE PERCEPTION

3.1 Procédure

Stimuli: 20 syllabes (fricative + voyelle) ont été générées à l'aide de Compost, système de synthèse par règles utilisant le synthétiseur de Klatt et développé par l'ICP de Grenoble (Ref.7). Les cibles formantiques ont été modifiées de telle sorte à générer un ensemble de syllabes ambiguës, dont l'identité de la consonne initiale varie entre [f,s,ch] et celle de la voyelle entre [i,é,y,oe,e,o,u].
Seules les valeurs des fréquences des 3 premiers formants ainsi que leurs intensités relatives et largeurs de bande étaient variables. Les valeurs choisies pour les fréquences formantiques sont fournies dans le Tab.2. Les 20 stimuli correspondent aux combinaisons de 4 cibles consonnantiques et 5 cibles vocaliques.

Tab.2 Fréquences des formants pour les stimuli synthétiques.

Cibles consonnantiques	Fric1	Fric2	Fric3	Fric4
F1	300	300	300	300
F2	1320	1500	1680	1860
F3	2250	2430	2610	2790

Cibles vocaliques	Voc1	Voc2	Voc3	Voc4	Voc5
F1	275	283	290	300	310
F2	2100	1750	1400	1050	700
F3	3100	2850	2600	2350	2100

Séries: une pré-série de 10 stimuli, servant à l'échauffement, ainsi qu'une série exérimentale comportant 5 exemplaires de chacun des 20 stimuli, en ordre aléatoire, ont été constituées et enregistrées sur lecteur de cassette analogique. La consigne était d'identifier aussi rapidement que possible, dans un délai maximal de 3 secondes, la syllabe en choisissant parmi l'une des réponses suivantes: "fi, fu, fé, fe, fo, feu, fou, si, su, sé, se, so, seu, sou, chi, chu, ché, che, cho, cheu, chou".
Sujets: 25 sujets francophones, sans troubles auditifs apparents, ont participé à l'expérience en étant rétribués à un taux fixe.

3.2 Résultats perceptifs

Le Tab.3 fournit les tendances globales de réponses, cumulées sur l'ensembles des stimuli. Les catégories de réponses [é, oe, o] sont regroupées avec les catégories de voyelles fermées correspondantes, respectivement [i, y, u]. L'intérêt de tableau est de fournir la relation entre les tendances de réponses vocaliques et consonnantiques. Les fréquences de réponses [ch], qui sont au total relativement rares, ne dépendent pratiquement pas du contexte vocalique. Les réponses [f] et [s] ont approximativement la même fréquence pour [i+é]. Par contre, les réponses [f] prédominent dans le contexte [y], tandis que les réponses [s] prédominent pour [u+o] et [e].

Tab.3 Réponses cumulées sur l'ensemble des 20 stimuli (* 25 sujets * 5 présent.).					
	i+é	y+oe	e	u+o	TOT
f	342	375	93	201	1011
s	354	266	156	350	1126
ch	123	112	57	71	363
TOT	819	753	306	622	2500

Les réponses dominantes pour les différentes combinaisons de formants fricatifs (FRI) et vocaliques (VOC) sont fournies dans le Tab.4.

Tab.4 Réponses modales pour les différents stimuli. Les zones hachurées signalent l'absence de mode dominant.					
	VOC1	VOC2	VOC3	VOC4	VOC5
FRI 1	fi .74	fi .46	fy .54	fy .53	fu .62
FRI 2	fi .61	fi .31	fy .44	fy .34	////
FRI 3	si .51	si .38	sy .35	////	su .74
FRI 4	si .70	si .39	sy .42	////	su .77

Afin de mettre en évidence les relations entre indices acoustiques et catégories de réponses, les résultats ont été traité par analyse multilogistique (Ref.9).

Suivant en celà la procédure utilisée par Nearey (Ref.3), trois modéles différents ont été appliqués: acoustique (AC), diphone additif (DA) et diphone interactif (DI). Dans le modèle AC seuls les valeurs des indices acoustiques sont utilisées pour prédire les résultats. Le modèle DA comporte des paramètres supplémentaires pour tenir compte des biais introduits par l'identité d'un segment sur l'identification du segment adjacent. Le modèle DI permet en outre de tenir compte de modifications éventuelles possibles dans le traitement des indices acoustiques dans l'identification d'un segment en fonction de l'identité du segment adjacent. Les performances des 3 modèles sont fournies dans le Tab.5.

Tab.5 Evaluation des modèles multilogistiques				
Model	X2	df	X2/df	F
AC	1410	205	6.877	1.74
DA	1033	199	5.192	1.32
DI	737	187	3.942	

Les tests X2 montrent que même pour le modèle saturé (DI) l'erreur résiduelle est significative et relativement élevée (près de 4 fois le nombre de degrés de liberté). Dans ces conditions, il est préférable de ne pas se fier aux valeurs du X2 pour tester les différences de performances entre modèles et une alternative simple consiste à se baser sur les rapports F. En procédant ainsi, la chute de performance pour le modéle AC, par rapport à DI, reste significative

($F_{205,187} = 1.74$; $p<.01$) et il en va de même pour l'écart entre les modèles DA et DI, mais avec un niveau de signification plus faible ($F_{199,187} = 1.32$; $p<.05$).

4. DISCUSSION GENERALE ET CONCLUSIONS

D'après les résultats obtenus ici, les relations entre le lieu d'articulation des fricatives et les voyelles adjacentes sont interactives, tant dans la production que dans la perception de syllabes CV en français. Ces interactions sont relativement complexes et ne peuvent pas encore être définies avec précision. Des analyses supplémentaires devraient permettre de les spécifier. Mais, par rapport à l'anglais, où des effets additifs ont été obtenus, l'adjonction de la voyelle arrondie d'avant [y] en français semble bien jouer un rôle, qui reste à préciser.

REFERENCES

(1) Carden, G., Levitt, A. Jusczyk, P.W. and Walley, A. (1981). Evidence for phonetic processing of cues to place of articulation: Perceived manner affects perceived place. *Perception and psychophysics* 29, 26-36.

(2) Serniclaes, W. and Wajskop, M. (in press) "Phonetic versus acoustic account of feature interaction in speech perception" to appear in *Analytic Approaches to Human Cognition, Proc. of the Conference in Honour of Paul Bertelson*, Brussels June 1991.

(3) Nearey, T.M. (1990) "The segment as a unit of speech perception" *J. of Phonetics* 18, 347-373.

(4) Nearey, T.M. (1991) Perception: Automatic and Cognitive *Proc. 12th Int.Cong. Int. Phonetic Sciences, Aix-en-Provence 1991*, Vol.1, 40-49.

(5) Serniclaes, W. (1991). "Perceptual processing and ecological validity" *Proc. 12th Int. Cong. Phonetic Sciences, Aix-en-Provence 1991*, Vol. 1, 50-55.

(6) Whalen, D.(1989) "Vowel and consonant judgments are not independent when cued by the same information" *Perception and Psychophysics*, 46, 284-292.

(7) Brunswik,E. (1957). Scope and aspects of the cognitive problem. In *Contemporary Approaches to Cognition.* Cambridge, MA: Harvard U. Press; 5-40.

(8) Bailly,G. and Guerti,M. (1991) "Synthesis-by-rule for French" *Proc. 12th Int. Cong. Phonetic Sciences, Aix-en-Provence 1991*, Vol. 2, 506-509.

(9) McCullagh,P. and Nelder,J.A. (1983) *Generalized Linear Models* London: Chapman & Hall.

19es J.E.P. - Bruxelles, 19 au 22 mai 1992.

REPRESENTATIONS INTERMEDIAIRES DANS LA RECONNAISSANCE DE LA PAROLE: APPORTS DE LA TECHNIQUE DE CREATION DE MOTS ILLUSOIRES

REGINE KOLINSKY*+ ET JOSE MORAIS*

*LABORATOIRE DE PSYCHOLOGIE EXPERIMENTALE
UNIVERSITE LIBRE DE BRUXELLES; + F.N.R.S.

Résumé

Le format des représentations intermédiaires de la parole a été étudié au travers de l'occurrence d'erreurs de migration. Ces erreurs ont été induites par la présentation de paires dichotiques dans lesquelles l'information nécessaire à la perception d'une cible était distribuée entre les deux stimuli. La distribution de l'information entre les stimuli a été manipulée afin de comparer les migrations de voisement, de lieu d'articulation, de consonne, de voyelle, et de syllabe.

Les résultats suggèrent que la voyelle, la syllabe et le voisement participent aux représentations intermédiaires. De plus, l'aptitude des différentes propriétés à migrer n'était pas affectée différentiellement par le statut lexical de la cible, ce qui suggère que le processus de segmentation sous-tendant le phénomène de migration est pré-lexical.

Les modèles psycholinguistiques envisagent la reconnaissance de la parole comme un appariement entre le signal acoustique et les représentations des mots stockées en mémoire, que l'on appelle représentations lexicales. Deux classes principales de modèles ont été proposés. Selon l'une, l'information extraite de l'entrée sensorielle est appariée directement au lexique, de manière continue, au travers du calcul soit de gabarits spectraux (Klatt, 1980), soit de traits phonétiques (Lahiri & Marslen-Wilson, 1991; Stevens, 1986). La seconde classe de modèles, plus populaire, suppose que certaines représentations sous-lexicales jouent le rôle de médiateurs dans le processus d'appariement entre le signal acoustique et le lexique mental. La nature de des représentations intermédiaires reste cependant inconnue. Deux unités linguistiques ont souvent été proposées: le phonème et la syllabe. D'autres candidats, tels que la demi-syllabe (Samuel, 1989), pourraient également être envisagés. Dans tous les cas, un seul niveau de représentations intermédiaires est proposé, représentations qui joueraient le rôle à la fois de sortie du traitement acoustique-phonétique et celui d'entrée au traitement lexical. Cette vue unitaire pourrait être trop simple: il est possible que les unités de segmentation de la parole soient différentes des unités de classification et d'accès qui établissent le contact avec le lexique (Cutler et Norris, 1988). Il est aussi possible que le système exploite différents types d'information, et les représentations sous-lexicales seraient en conséquence hautement structurées (Church, 1987; Frazier, 1987).

Un problème majeur dans l'étude du format des représentations intermédiaires est que l'utilisation de techniques de détection d'unités sous-lexicales (Mehler, Dommergues, Frauenfelder et Segui, 1981; Savin et Bever, 1970) ne permet pas de séparer les représentations intermédiaires des représentations ultérieures, résultantes de l'atteinte du lexique (Frauenfelder, 1991; Morais, 1985).

Dans la nouvelle situation de détection que nous présentons, à la fois les stimuli et les cibles étaient des mots ou des pseudo-mots, la tâche consistant à détecter une cible parmi une paire de stimuli dichotiques. La situation intéressante est celle où l'auditeur percevait illusoirement une cible absente. L'occurrence de telles erreurs était induite par la présentation de paires dans lesquelles l'information nécessaire à la perception de la cible (par exemple, "BIJOU", /b i ʒ u/) était distribuée entre les deux stimuli (par exemple, "KIJOU-BOTON", /k i ʒ u/ -/b ɔ t õ/ pour la première consonne, C_1). La combinaison de parties d'information d'une représentation d'entrée avec des parties temporellement contigües de l'autre représentation créait l'illusion. Le taux de migrations observé reflèterait l'aptitude de l'unité sous-lexicale envisagée à être représentée en tant qu'unité séparée dans les représentations intermédiaires de la parole.

La distribution de l'information entre lès deux stimuli dichotiques a été manipulée afin de comparer directement différentes unités sous-lexicales. La même

cible, par exemple /b i ʒ u/, pouvait être créée illusoirement à partir d'une migration de voisement de C_1 (stimuli: /p i ʒ u/-/g ɔ t̃/), de lieu d'articulation de C_1 (/g i ʒ u/-/p ɔ t̃/), de C_1 (/k i ʒ u/-/b ɔ t̃/) de première voyelle, V_1 (/b ɔ ʒ u/-/k i t̃/) ou de syllabe (/b i t̃/-/k ɔ ʒ u/). Afin de vérifier que l'illusion ne résultait pas de confusions perceptives, des paires contrôles dans lesquelles un des stimuli ne présentait pas l'unité critique ont été utilisées. Les paires contrôles contenaient toujours un des pseudo-mots de la paire expérimentale correpondante, à savoir le plus proche au niveau perceptif de la cible; l'autre pseudo-mot était changé de manière minimale. Par exemple, étant donné la cible /b i ʒ u/, dans la condition C_1 l'essai expérimental était /k i ʒ u/ - /b ɔ t̃/; dans l'essai contrôle, le pseudo-mot /k i ʒ u/ était maintenu, et le pseudo-mot /d ɔ t̃/, qui ne présente pas la valeur labiale, remplaçait /b ɔ t̃/. Pour la condition syllable, étant donné qu'il était arbitraire de décider à priori lequel des deux stimuli expérimentaux était le plus proche de la cible, nous avons utilisé deux types différents d'essais contrôles, partageant avec la cible soit la syllable initiale (comme dans /b i t̃/ - /k ɔ ʃ o/ pour la cible /b i ʒ u/), soit la syllabe finale (comme dans /d e t̃/-/k ɔ ʒ u/).

Pour la session principale, les paires expérimentales et contrôles négatives, dans lesquelles la cible était absente, ont été créées sur base de 3 ensembles de paires de pseudo-mots qui, par échange d'une des 5 propriétés linguistiques, pouvaient produire un mot cible. Les mots-cibles étaient, pour un ensemble, BIJOU et COTON (/b i ʒ u/, /k ɔ t̃/), pour un autre ensemble DELIT et PINCEAU (/d e l i/, /p ɛ̃ s o/), et pour le troisième TISSU et GAMIN (/t i s y/, /g a m ɛ̃/). Les paires où la cible était un pseudo-mot ont été construites en intervertissant la deuxième consonne des stimuli et des deux cibles d'un ensemble. Ceci donnait les cibles /b i t u/ et /k ɔ ʒ õ/ pour le premier ensemble, /d e s i/e t /p ɛ̃ l o/ pour le deuxième, et /t i m y/ e t /g a s ɛ̃/ pour le troisième. Pour chaque cible, il y avait 11 types différents d'essais négatifs et, de plus, 15 types différents de paires positives, dans lesquelles la cible était présente.

Tous les stimuli étaient des pseudo-mots CVCV obéissant à la phonotactique du français. Les consonnes initiales étaient toujours des plosives, et celles-ci présentaient toujours un double contraste (voisement et lieu d'articulation) dans les paires expérimentales négatives. Chaque type d'essai négatif était présenté six fois, trois fois à une oreille et trois fois à l'autre. Chaque type d'essais positif était présenté deux fois, avec aussi un balancement inter-oreilles.

Les stimuli, prononcés par un francophone natif, ont été enregistrés dans une chambre anéchoïque, et ont ensuite été digitalisés et traités sur une station de travail informatisée. Les stimuli d'une paire dichotique ont été normalisés et synchronisés au début et à la fin des items, ainsi qu'au début de la seconde syllabe. Ils étaient présentés par l'intermédiaire d'écouteurs. Les sujets, 20 étudiants universitaires de langue maternelle française, devaient lire la cible dans un carnet. On leur demandait de faire attention aux messages provenant des deux oreilles afin de décider, à chaque essai, si oui ou non la cible était présente.

Les analyses effectuées sur les résultats portent sur le paramètre de la théorie de la détection, *d'* (Green et Swets, 1966). Ce paramètre mesure la discriminabilité entre la situation où la cible est présente et celle où la cible est absente, indépendamment des biais de réponse éventuels des sujets. Nous l'avons calculé pour chaque sujet, séparément pour les essais expérimentaux (E) et pour les essais contrôles (C), pour chaque type de propriété linguistique testé, et pour les cibles mots et pseudo-mots. Nous avons ensuite calculé des taux de migration sur ces valeurs de *d'*, pour chaque sujet et pour chaque type de propriété, séparément pour les cibles mots et pseudo-mots. Le taux de migration est équivalent à la proportion: *d'* aux essais expérimentaux/ somme des *d'* expérimentaux et des *d'* contrôles. Les moyennes des taux de migration sont présentées à la Table 1. Sous l'hypothèse que les propriétés des représentations intermédiaires de la parole migrent, on peut dériver la prédiction que les essais E constituent une situation de moindre discriminabilité que les essais C (Treisman et Souther, 1986). Les scores *d'* devraient dès lors être moins élevés pour les essais E que pour les essais C. Ceci implique que les taux calculés à partir des scores *d'* sont d'autant plus petits, par rapport à 0.50, qu'il y a plus de migrations. Les valeurs des tests t concernant cette prédiction, ainsi que les niveaux de signification, sont présentés à la Table 1, séparément pour les cibles mots et les cibles pseudo-mots.

Cette première analyse montre que la voyelle, la syllabe, et, marginalement, le voisement, semblent participer aux représentations intermédiaires de la parole. Afin de pouvoir comparer les différents types de propriétés entre elles, nous avons effectué une analyse de variance, qui, pour éviter d'éventuels

problèmes d'échelle de mesure, prend en compte les transformées arcsinus des taux de migration calculés ur *d'*. Les scores ainsi obtenus sont visibles à la Figure 1.

TABLE 1: Taux de migration, valeur et signification (P) des tests t

	VOISEMENT	LIEU	C1	V1	SYLLABE
CIBLES MOTS					
TAUX sur FDs					
X	.64	.49	.58	.61	.80
Sx	(.09)	(.03)	(.07)	(.13)	(.14)
TAUX sur d'					
X	**.38**	**.48**	**.49**	**.37**	**.33**
Sx	(.16)	(.35)	(.13)	(.13)	(.14)
t(19)=	3.19	.28	.37	4.76	5.42
P≤	**.003**	**.39**	**.36**	**.0001**	**.0001**
CIBLES PSEUDO-MOTS					
TAUX sur FDs					
X	.64	.47	.51	.58	.71
Sx	(.14)	(.05)	(.08)	(.11)	(.16)
TAUX sur d'					
X	**.43**	**.57**	**.63**	**.43**	**.48**
Sx	(.17)	(.25)	(.14)	(.09)	(.10)
t(19)=	1.8	—	—	3.28	0.85
P≤	**.04**			**.002**	**.20**

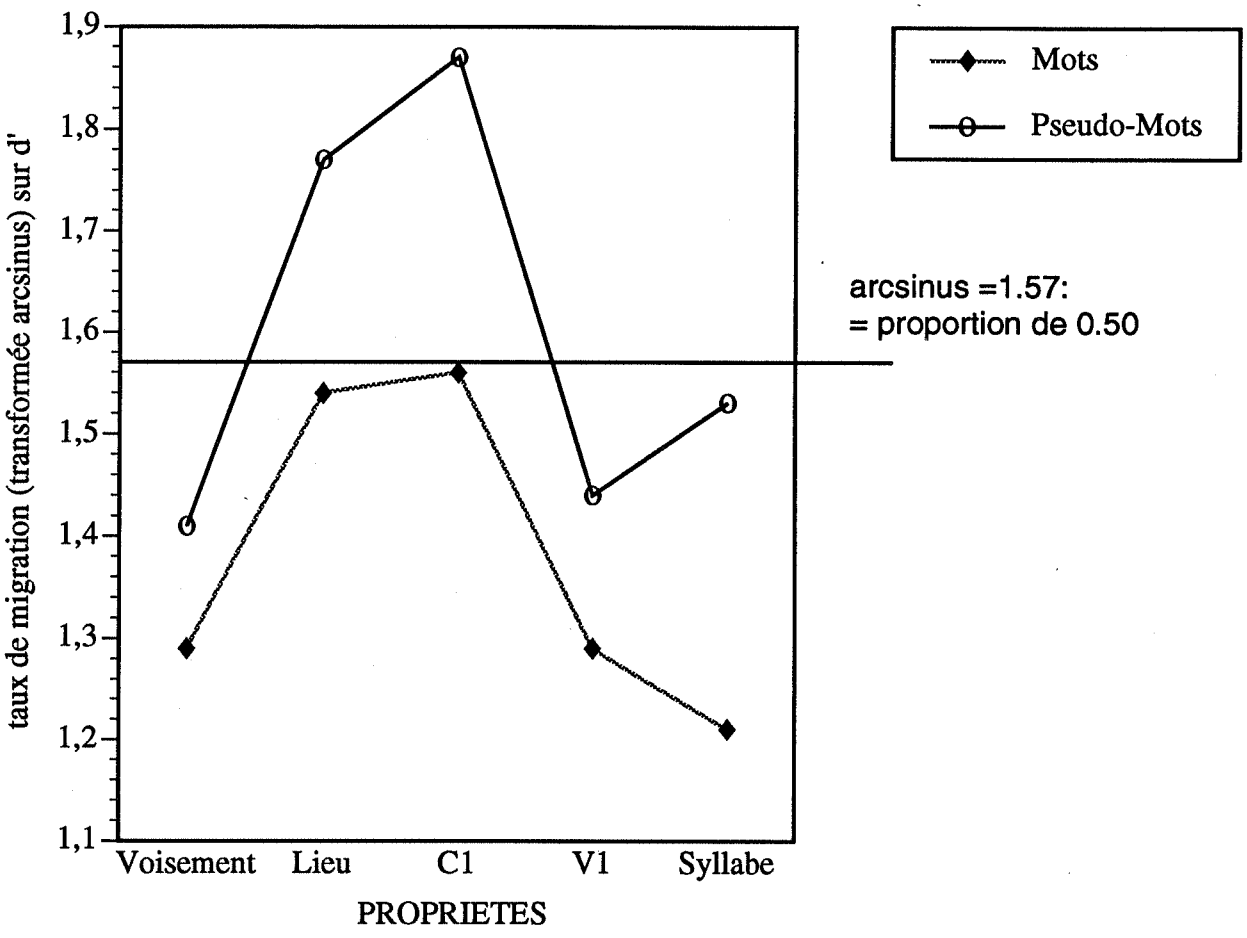
L'analyse, qui tient compte des cinq types de propriétés testés ainsi que du statut lexical ou non de la cible, montre des effets hautement significatifs du type de propriété ($F_{4-76} = 4.79$, $P < .005$) ainsi que de la lexicalité de la cible ($F_{1-19} = 12.63$, $P < .005$). Cependant, l'interaction entre ces deux facteurs n'est pas significative ($F < 1$). Les comparaisons deux à deux entre types de propriété montrent de plus que les migrations de syllabe et de V_1, dont les fréquences d'occurrence sont comparables ($F < 1$), sont significativement plus fréquentes que les migrations de C_1 ($F_{1-19} = 35.58$, $P <. 0005$ et $F_{1-19} = 22.38$, $P < .0005$, respectivement).

Dans la mesure où l'occurrence de migrations de syllabe est particulièrement importante, on peut se demander si ce que nous avons défini comme étant une migration de nature syllabique ne résulte pas simplement de migrations indépendantes mais simultanées de C_1 et de V_1. Nous avons donc testé, indépendamment pour les cibles mots et pour les cibles pseudo-mots, si la syllabe migre réellement en tant qu'unité globale. Les comparaisons des probabilités de migration calculées sur les pourcentages de fausses détections (FDs), et qui sont présentées dans la Table 1, ont montré que la probabilité des migrations de syllabes était significativement supérieure au produit des probabilités des migrations de C_1 et de V_1 ($t_{19} = 14.09$ et $= 11.3$, $P < .0001$, pour les cibles mots et pseudo-mots, respectivement). Il semble donc que les syllabes migrent bien de manière globale.

En conclusion, le fait que l'aptitude des différentes propriétés à migrer n'était pas affectée différentiellement par le statut lexical de la cible suggère que le processus de segmentation sous-tendant le phénomène de migration est pré-lexical. L'ensemble

des résultats suggère donc que, contrairement aux consonnes, à la fois le voisement, les voyelles et les syllabes pourraient jouer un rôle de médiateur dans le processus d'appariement entre le signal acoustique et le lexique mental. Ces résultats s'opposent à l'idée que seuls les traits phonétiques seraient extraits, en continu, du courant acoustique, de même qu'à l'idée qu'on calculerait seulement des représentations spectrales temporellement délimitées. Au contraire, il existerait des représentations sous-lexicales intermédiaires dans le processus d'appariement entre le signal acoustique et le lexique mental. De plus, nos résultats montrent qu'il existe des représentations sous-lexicales multiples.

FIGURE 1



Cependant, les résultats rapportés ici pourraient n'avoir qu'une portée limitée au français: les processus de segmentation de la parole pourraient dépendre fortement des caractéristiques phonologiques de la langue des auditeurs (Cutler, Mehler, Norris et Segui, 1986). En ce sens, le fait que nous observons dans la présente expérience une prépondérance de migrations de voyelles et de syllabes par rapport aux consonnes n'est pas étonnant: en français, les frontières syllabiques sont relativement claires, la syllabe est l'unité rythmique de base, et les voyelles ne sont que rarement réduites. Mais l'importance relative des phonèmes et des syllabes pourrait varier par rapport au français dans des langues où l'unité de rythme est différente (telle que par exemple la mora, en japonais), ou dans des langues où les patrons accentuels sont différents (comme c'est le cas des langues à accent variable, telles que l'anglais), ou encore dans des langues où le phénomène de réduction vocalique induit de longues suites consonantiques (comme par exemple en portugais). L'influence de chacun de ces facteurs devra être examinée par les comparaisons inter-langues appropriées. Une première série de résultats, portant sur l'influence du phénomène

de réduction vocalique, soutient clairement la notion d'une dépendance phonologique de la nature des unités intermédiaires dans la reconnaissance des mots parlés.

REFERENCES

Church, K. (1987) Phonological parsing and lexical retrieval. *Cognition*, **25**, 53-69.

Cutler, A., et Norris, D. (1988). The role of strong syllables in segmentation for lexical access. *Journal of Experimental Psychology: Human Perception and Performance*, **14**, 113-121.

Cutler, A. Mehler, J. Norris, D. et Segui, J. (1986) The syllable's differing role in the segmentation of English and French. *Journal of Memory and Language*, **25**, 385-400.

Frauenfelder, U.H. (1991) The interface between acoustic-phonetic and lexical processing. Papier presenté à l'*OTS Workshop on the Psychophysics of speech perception II*, Utrecht.

Frazier, L. (1987) Structure in auditory word recognition. *Cognition*, **25**, 157-187.

Green, D. M., et Swets, J. A. (1966) *Signal detection theory and psychophysics*. New York: Wiley.

Mehler, J. Dommergues, J. Y., Frauenfelder, U. H., et Segui , J. (1981) The syllable's role in speech segmentation. *Journal of Verbal Learning and Verbal Behavior* **20**, 298-305.

Morais, J. (1985) Literacy and awareness of the units of speech: Implications for research on the units of perception. *Linguistics* **23**, 707-721.

Klatt, D.H. (1980) Speech perception: A model of acoustic-phonetic analysis and lexical access. Dans R.A. Cole (Ed.) *Perception and production of fluent speech.* N.J.: Lawrence Erlbaum Ass.

Lahiri, A. et Marslen-Wilson, W. D. (1991) The mental representation of lexical form: phonological approach to the recognition lexicon. *Cognition*, **38**, 245-294.

Samuel, A.G. (1989) Insights from a failure of selective adaptation: Syllable-initial and syllable-final consonants are different. *Perception and Psychophysics*, **45**, 485-493.

Savin, H. B. et Bever, T. G. (1970) The non-perceptual reality of the phoneme. *Journal of Verbal Learning and Verbal Behavior* **9**, 295-302.

Stevens, K.N. (1986) Models of phonetic recognition II: An aaproach to feature-based recognition. Dans P. Mermelstein (Ed.) *Proceedings of the Montreal Satellite Symposium on Speech Recognition, Twelfth International Congress on Acoustics.*

Treisman, A. et Souther, J. (1986) Illusory words: the roles of attention and of top-down constraints in conjoining letters to form words. *Journal of Experimental Psychology: Human Perception and Performance*, **12**, 3-17.

REMERCIEMENTS

Ce travail fait partie d'un projet internatinal intitulé *Processing consequences of contrasting language phonologies* et subsidié par le Human Frontier Science Program. Des subsides ont également été accordés par le Fonds National de la Recherche Scientifique (F.N.R.S.)- Loterie Nationale (convention n° 8.4527.90) et par le Ministère de l'Education de la Communauté française de Belgique (Action de Recherche concertée: *Le traitement du langage dans différentes modalités: approches comparatives.* Le premier auteur est Chercheur Qualifié du F.N.R.S. Nous remercions spécialement M. Cluytens pour la préparation du matériel, N. Martin pour le traitement des données et l'Institut de Phonétique de l''U.L.B. pour les facilités d'enregistrement.

19es J.E.P. - Bruxelles, 19 au 22 mai 1992.

LE ROLE DU VOISINAGE DANS L'AMORCAGE PHONOLOGIQUE

MONIQUE RADEAU & JOSE MORAIS

UNIVERSITE LIBRE DE BRUXELLES
LABORATOIRE DE PSYCHOLOGIE EXPERIMENTALE

Résumé

L'amorçage phonologique auditif a été examiné avec un plan expérimental permettant de vérifier d'éventuelles interactions entre quatre facteurs: type de recouvrement phonologique entre l'amorce et la cible (initial/final), intervalle inter-stimuli (20/500 msec), fréquence relative (soit amorce de basse fréquence et cible de haute fréquence, soit l'inverse), et tâche (décision lexicale/répétition). Le résultat principal est un effet systématique de facilitation du recouvrement final, contrastant avec l'absence d'effet du recouvrement initial.

L'étude de la reconnaissance des mots parlés par l'observation d'effets d'amorçage phonologique a fait l'objet, ces dernières années, d'un certain nombre de tentatives. L'idée de base est que, quand l'amorce et la cible partagent certaines propriétés phonologiques, il peut y avoir soit activation soit inhibition de la cible. L'hypothèse sur laquelle repose cette idée est que l'activation des entrées lexicales à partir des signaux de parole passe par la constitution de représentations des composantes phonétiques ou phonologiques (traits, phonèmes, attaques, rimes, syllabes, etc...).

Malheureusement, les résultats des études publiées sont des plus contradictoires. Ainsi, Slowiaczek, Nusbaum & Pisoni (1987) ont obtenu un effet de facilitation dans une tâche d'identification de mots masqués par du bruit blanc quand l'amorce et la cible ont en commun deux phonèmes initiaux au moins. Pourtant, ni Slowiaczek & Pisoni (1986), utilisant le même matériel que dans l'étude précédente, ni Jakimik, Cole & Rudnicky (1985) n'ont trouvé d'effet d'amorçage phonologique dans la tâche de décision lexicale.

En outre, alors que dans la tâche de décision lexicale, Radeau, Morais & Dewier (1989) ont obtenu un effet d'interférence en cas d'identité d'un ou de deux phonèmes initiaux entre l'amorce et la cible, Emmorey (1989) a obtenu un effet de facilitation entre items qui soit rimaient soit partageaient la syllabe finale non-accentuée. L'incohérence apparente de ces résultats pourrait être levée si la position initiale versus finale du recouvrement phonologique entre l'amorce et la cible était un facteur critique de la nature de l'effet.

Avec une autre tâche, celle de répétition immédiate de la cible, Radeau et al. (1989) n'ont trouvé aucun effet d'amorçage phonologique dans des conditions permettant d'éviter des stratégies de prédiction du mot-cible. Par contre, avec cette même tâche de répétition, Slowiaczek & Hamburger (in press) ont trouvé un effet facilitateur en cas d'identité du phonème initial et un effet d'interférence en cas d'identité de deux ou plusieurs phonèmes initiaux. Il est difficile de concilier cet ensemble de résultats avec ceux obtenus en décision lexicale, mais il faut remarquer qu'aucune étude publiée jusqu'ici n'a croisé le facteur tâche (décision lexicale versus répétition immédiate) avec la position (initiale versus finale) du recouvrement phonologique.

La recherche présentée ici avait cet objectif. Nous pensions qu'une étude systématique des variables qui peuvent affecter l'effet d'amorçage phonologique est nécessaire si l'on veut utiliser cet effet comme un instrument d'examen des théories de la reconnaissance des mots parlés.

Notre matériel était composé uniquement d'items monsyllabiques (en grande majorité de CVC), et nous avons manipulé, outre la position du recouvrement (tous les phonèmes sauf le dernier versus tous les phonèmes sauf le premier) et la tâche (décision lexicale versus répétition immédiate), deux autres facteurs : la durée de l'intervalle entre l'amorce et la cible et la fréquence relative de l'amorce et de la cible.

L'intervalle temporel peut être de grande importance pour l'observation d'effets d'amorçage phonologique, puisque l'activation et l'inhibition des entrées lexicales pourraient être des phénomènes hautement transitoires. Nous avons donc utilisé deux intervalles -de 20 et 500 msec respectivement- entre la fin de l'amorce et le début de la présentation de la cible. La fréquence relative quant à elle a été introduite dans ce travail suite à l'observation d'effets de cette variable dans l'amorçage de la reconnaissance de mots écrits par des "voisins" phonologiques ou orthographiques (Colombo, 1986; Grainger, 1990; Segui & Grainger, 1990).

Dans notre étude, nous avons utilisé des paires amorce-cible telles qu'un des items était de haute fréquence et l'autre de basse fréquence. L'effet de la fréquence relative a été testé en comparant dans chaque expérience deux groupes différents de sujets. Un groupe recevait les amorces de basse fréquence suivies des cibles de haute fréquence, et l'autre groupe était exposé à la combinaison inverse. Nous avons réalisé quatre expériences résultant chacune du croisement des deux intervalles interstimuli avec les deux tâches. Dans chaque expérience, les deux conditions de recouvrement ou, en d'autres termes les deux conditions de voisinage phonologique, étaient considérées. Pour chacune de ces conditions il y avait 8 essais liés (avec lien phonologique) et 8 essais contrôles (sans lien phonologique).

Les mots ont été enregistrés par un locuteur masculin dans une chambre sourde, puis digitalisés avec Sound Tools (16 bits, 32000 Hz). Ils ont été stockés par paires sur le canal 1 d'un enregistreur DAT. Un clic, inaudible pour le sujet et servant à déclencher la carte clock d'un ordinateur APPLE 2E a été enregistré sur le canal 2.

Les résultats des analyses menées sur les TR moyens (calculés séparément pour chacun des sujets et chacun des items) de chacune des quatre expériences ont mis en évidence une interaction significative entre le type d'essai (lié/non-lié) et le type de recouvrement phonologique (initial/final). Cette interaction reflète le fait que, contrairement à la condition de recouvrement phonologique final qui donne lieu à un effet systématique de facilitation, la condition de recouvrement initial ne donne lieu à aucun effet d'amorçage. Notons que cette interaction était également significative dans l'analyse de variance regroupant les résultats des quatre expériences.

Au niveau de l'analyse globale, nous n'avons pas obtenu d'interaction qui soit significative à la fois dans l'analyse par sujet et l'analyse par item entre la variable essai et les autres variables envisagées. Toutefois, certaines interactions significatives à l'une de ces analyses seulement valent la peine d'être prises en compte.

Ainsi, l'interaction, entre essai, intervalle temporel et type de recouvrement phonologique était significative à l'analyse par sujet mais pas à l'analyse par item, indiquant une tendance pour l'effet d'amorçage obtenu dans la condition de recouvrement final à diminuer quand l'intervalle temporel entre amorce et cible augmente. Il est donc possible qu'avec un intervalle plus long, les effets d'amorçage diminuent significativement et même disparaissent. En outre, dans l'analyse portant sur les résultats de la condition d'intervalle temporel court (20 msec), l'interaction entre essai, fréquence relative et recouvrement phonologique était également significative par sujet mais pas par item. Cette interaction est indicative d'une tendance pour l'effet de facilitation obtenu dans la condition de recouvrement phonologique final à être moins important quand l'amorce est de haute fréquence et la cible de basse fréquence que dans la condition de fréquence relative opposée.

Le résultat majeur de cette étude est que l'amorçage phonologique dépend du type de recouvrement qui existe entre l'amorce et la cible. Dans les essais avec recouvrement final, la relation phonologique était une relation de rime. Une interprétation vraisemblable de ce résultat est que l'activation d'une forme phonologique conduise à l'activation des formes phonologiques qui riment avec la première. Les données actuelles ne permettent pas de savoir si ces formes phonologiques sont limitées au lexique. Nous envisageons donc de faire des expériences similaires où la cible est un non-mot. Nous comptons aussi vérifier si le même type d'effet apparaît en japonais, langue où la rime n'est pas un instrument de manipulation linguistique aussi important qu'en français.

Par ailleurs, l'absence totale d'effet d'amorçage en cas de partage des deux phonèmes initiaux suggère que, au moins pour des mots courts monosyllabiques, la reconnaissance des mots en français ne passerait pas par la constitution de représentations des phonèmes, ou que

si ces représentations sont constituées, leur activation s'éteint extrêmement vite.

Références

Columbo, L. (1986). Activation and inhibition with orthographically similar words. *Journal of Experimental Psychology: Human Perception and Performance, 12*, 226-234.

Emmorey K.D. (1989). Auditory morphological priming in the lexicon. *Language and Cognitive Processes, 4*, 73-92.

Grainger, J. (1990). Word frequency and neighborhood frequency effects in lexical decision and naming. *Journal of Memory and Language, 29*, 228-244.

Jakimik, J., Cole, R.A. & Rudnicky, A.I. (1985). Sound and spelling in spoken word recognition. *Journal of Memory and Language, 24*, 165-178.

Radeau, M., Morais, J. & Dewier, A. (1989). Phonological priming in spoken word recognition: Task effects. *Memory and Cognition, 17*, 525-535.

Segui, J. & Grainger, J. (1990). Priming word recognition with orthographic neighbors: Effects of relative prime-target frequency. *Journal of Experimental Psychology: Human Perception and Performance, 16*, 65-76.

Slowiaczek, L.M. & Hamburger, M. (in press). Pre-lexical facilitation and lexical interference in auditory word recognition. *Journal of Experimental Psychology: Learning, Memory and Perception.*

Slowiaczek, L.M., Nusbaum, H.C. & Pisoni, D.B. (1987). Phonological priming in auditory word recognition. *Journal of Experimental Psychology: Learning, Memory and Cognition, 13*, 64-75.

Slowiaczek, L.M. & Pisoni, D.B. (1986). Effects of phonological similarity on priming in auditory lexical decision. *Memory and Cognition, 14*, 230-237.

19es J.E.P. - Bruxelles, 19 au 22 mai 1992.

LES INTERFERENCES PHONETIQUES COMME APPROCHE DES CONFUSIONS PERCEPTIVES.

SOPHIE WAUQUIER-GRAVELINES

LABORATOIRE DE PHONETIQUE DU D.R.L PARIS 7

Résumé

En 1955, Miller et Nicely ont proposé d'étudier les confusions perceptives en établissant des matrices.Nous proposons une étude on-line des confusions perceptives par le recours au phénomène des interférences phonologiques,mentionné pour la première fois par Newman et Dell (78). Utilisant la tâche de détection de phonèmes les sujets ont détecté moins rapidement par ex. /b/ dans "petit balai" que dans "vrai balai".Les résultats suggèrent que le nombre et la nature des traits influencent le risque de confusion.

INTRODUCTION

Il est rare que la parole humaine soit perçue dans des conditions optimales. Face aux erreurs de production provenant de la source d'émission existent également des erreurs de perception. Les phénomènes de confusion perceptive ont à ce jour été quelque peu négligés par la recherche en Psycholinguistique alors qu'ils peuvent permettre d'élaborer des hypothèses éclairant les modèles de perception de la parole.

L'identification des signes linguistiques et plus particulièrement des mots, dépend de la capacité qu'a l'auditeur d'utiliser le système phonologique de sa langue et notamment de discriminer les phonèmes et de les opposer Il faut pouvoir opposer /p/ et /b/ pour pouvoir discriminer "pont" et "bon". Le problème posé ici ne consiste pas cependant à tester la pertinence psychologique de tel ou tel système de traits élaborés par différentes écoles linguistiques.Il ne s'agit pas d'infirmer ou de confirmer une représentation formalisée fondée sur des différences acoustiques ou articulatoires, mais d'étudier les oppositions entre les sons d'une langue d'un point de vue perceptif, en se centrant non pas sur leurs valeurs distinctives mais sur la manière dont la nature de ces oppositions influence la perception.

En 1955, Miller et Nicely ont abordé ce point en étudiant les confusions perceptives existant entre 16 phones consonnantiques de l'anglais. Les 16 consonnes-tests étaient prononcées en début de syllabe devant la voyelle a, et le sujet devait après chaque écoute

répéter ce qu'il avait perçu. Six rapports signal/bruit ont été utilisés.L'analyse des matrices de confusion permet d' observer que la plupart des confusions ont lieu entre des phones qui partagent un grand nombre de traits.Toutefois l'importance respective des traits est très variable : Selon leurs résultats, les confusions entre les consonnes occlusives sont plus fréquentes entre celles qui diffèrent par le lieu d'articulation qu'entre celles qui diffèrent par le voisement.
On peut cependant objecter que ces résultats ont été obtenus d'une part avec des tâches off-line et, d'autre part, en utilisant du signal bruité.
Nous proposons, par le recours au phénomène des interférences une alternative à cette étude des confusions perceptives. Le phénomène a été repéré par Newman et Dell(78), lors d'une recherche sur les ambiguités lexicales utilisant le paradigme de la détection de phone. Le temps de détection d'un phone-cible /b/ placé au début d'un mot de la phrase est plus long si le phone initial du mot précédent partage de nombreux traits avec la cible. Ainsi le TR pour la détection du /b/ dans "beach" est plus long s'il est précédé par le mot "private" que s'il est précédé par le mot"secret". La distance entre deux phones exprimée en nombre de traits partagés, est étroitement corrélée avec leur ressemblance perceptive et donc avec la possibilité de les confondre.En outre cette procédure expérimentale utilise la tâche de détection de phone qui permet d'observer le comportement perceptif des sujets "on-line" et sans s'appuyer sur leurs jugements métalinguistiques.
Il nous a donc paru intéressant d'utiliser ce phénomène comme paradigme pour l'étude de l'identification et de la discrimination de sons de parole non déformés en temps réel.
L'expérience présentée ici a été précédée de plusieurs expériences pilotes: nous voulions en effet tenter d'établir si le phénomène pouvait être obtenu sur listes de mots isolés alors qu'il n'avait été testé qu'en contexte phrastique. L'utilisation de ce type de matériel permet une manipulation expérimentale plus aisée des mots comportant des phones susceptibles d'être confondus, tout en réduisant l'information éventuelle des niveaux supralexicaux de traitement.Toutes ces expériences se sont soldées par des échecs: malgré la manipulation des intervalles séparant les mots, de la longueur des listes, du débit de lecture, les mots étaient traités séparément. Ceci supposerait que le phénomène d'interférence ne se produit que sur du matériel proche de la parole continue sans que l'on sache pourtant quels facteurs-prosodie, coarticulation, relation syntaxique entre les mots- sont déterminants.
Nous avons donc opté pour des listes d'énoncés courts (ex un beau palais),réalisant ainsi un compromis entre mots et phrases. Ceci nous permettait de maintenir les caractéristiques de la parole continue tout en gardant la souplesse de manipulation du matériel.

EXPERIENCE

Dans cette expérience nous avons tenté d'établir si la distance phonologique entre deux sons influence le risque de les confondre.La distance entre deux sons est définie par le nombre et la nature des traits qui les opposent.La définition des traits à laquelle nous nous référons, renvoie aux caractéristiques articulatoires des sons (voisement,lieu d'articulation type d'articulation).
Nous proposons que l'intensité de l'interférence entre les deux sons exprimée par la différence des TR aux différents items constitue une mesure d'évaluation du risque de confusion perceptive entre ces deux sons.Nous avons choisi comme phonème cible les deux occlusives bilabiales /p/ et /b/ que nous avons présentées dans les conditions suivantes:
-phone critique séparé par un seul trait :voisement
ex un beau palais
un petit balai

phone critique séparé par deux traits:voisement et place d'articulation.
ex un doux parfum
le tigre bondit
Nous souhaitions observer comment la variation de la distance phonologique entre deux sons, exprimée en quantité de traits (1 ou 2) et en qualité de traits (voisement et place d'articulation) influence le degré de confusion entre ces deux sons.
I) Méthode
I-1 stimulis
Le matériel est constitué de deux listes contrebalancées d'énoncés courts (les items tests dans la liste 1 sont contrôles dans la liste 2 et vice-versa). Les énoncés sont, soit des syntagmes nominaux (ex le beau palais) soit de courtes phrases (ex le père boit). Les listes sont elles-mêmes constituées de 4 blocs présentés dans l'ordre suivant:
bloc 1 /p/ cible, /b/ critique 1 trait: voisement
bloc 2 /p/ cible, /d/ critique 2 traits: voisement et place
bloc 3 /b/ cible, /p/ critique 1 trait: voisement
bloc 4 /b/ cible, /t/ critique 2 traits: voisement et place
Nous avons choisi successivement comme cible les occlusives bilabiales sourde et voisée afin que le comportement des sujets ne soit pas imputable à la nature même du phonème à détecter. Les fillers sont des énoncés courts comportant la cible sur le premier ou le deuxième mot (ex il pleuvait hier, pardon Madame).Le nombre de fillers entre deux items testés est aléatoire et varie entre 2 et 5. Les listes ont été lues par un locuteur de langue maternelle française puis digitalisées afin de poser avant chaque mot-test des clics permettant d'enregistrer les temps de réponse des sujets.Les listes ont ensuite été recopiées sur une bande comportant le signal sur le premier canal et les clics sur l'autre.
I-2 procédure
Chacune des deux listes a été présentée à un groupe de 15 sujets. Nous avons utilisé la tâche de détection de phone qui permet d'observer le comportement automatique des sujets: Ceux-ci recevant la liste dans un casque à l'intérieur d'une cabine isolée devaient appuyer sur un boitier relié au fréquencemètre dès qu'ils entendaient un mot commençant par le

phone-cible qui leur avaient été désigné.Les clics placés au début de chaque mot test enclenchent un fréquence mètre et le signal envoyé par le sujet le stoppe.Le fréquence-mètre affiche donc le temps entre le début du mot et le moment où le sujet reconnaît le phone-cible et appuie sur le boîtier. Les temps de réponse ont été relevés manuellement et porté sur des feuilles réponses par l'expérimentateur. La consigne était donnée oralement avant la passation.

I-3 sujets

30 étudiants de la faculté R. Descartes en deuxième cycle ont passé ce test en deux groupes (15+15).Tous étaient de langue maternelle française et aucun d'entre eux ne souffrait de troubles de l'audition. Nous avons éliminé les temps qui pour chaque sujet étaient à plus de deux écarts-types de la moyenne des items tests d'une part et des items contrôles d'autre part.

II)Résultats

Les résultats par sujets sont donnés dans le tableau suivant. Ils indiquent les temps de réponse en ms à la tâche de détection de phonème.

	No1	No2
D1	580	539
D2	471	534
	109	6

Nous avions testé deux facteurs principaux:

1) Nombre de traits

No1:1 trait,voisement.

N2:2traits, voisement et place d'articulation

2) Distance phonologique

D1= items tests,

D2= items contrôles.

La différence entre les items tests et les items contrôles est très élevée (109 ms) quand ils ne sont séparés que par le seul trait de voisement; alors qu'elle est minime quand ils sont séparés par deux traits: voisement et place d'articulation.L'analyse de variance montre que le facteur D a un effet très significatif: $F(1,28)=39.28$ $p < .00001$, tandis que le facteur N n'introduit aucun effet. L'interaction entre les deux facteurs est également très significative $f(1,28)=35.1$ $p < .00001$.

Ces résultats traduisent le fait que l'effet d'interférence -et donc le risque de confusion perceptive- ne se produit que quand les items cibles et critiques ne sont séparés que par le seul trait de voisement.

Nous avons en outre constaté que ces résultats très significatifs n'étaient pas liés à la nature de la cible.

Le tableau suivant indique les résultats en fonction de la nature de la cible en condition No1 (/b/ cible et /p/ critique et /p/ cible et /b/ critique.)

	cibles	
	p	b
D1	594	558
D2	483	464
	111	93

On peut remarquer qu'il n'y a pas d'assymétrie significative

entre les deux situations et que l'opposition est constante, que la cible soit voisée ou non. Ceci laisserait donc penser que les sujets n'utilisent pas la différence acoustique entre /p/ et /b/ pour faciliter la détection du phone-cible.

DISCUSSION

Dans cette expérience où nous utilisons le phénomène d'interférence phonétique,nous proposons que l'intensité du phénomène d'interférence constitue la mesure d'évaluation du degré de confusion perceptive.
Les résultats confirment partiellement ceux obtenus par Miller et Nicely (55) avec d'autres procédures méthodologiques. Les sons-cibles qui ne diffèrent du son-critique que par le seul trait de voisement sont détectés plus lentement que ceux qui diffèrent par deux traits. Plus les phones partagent de traits, plus ils sont susceptibles d'être confondus. La question suivante se pose néammoins : le degré de confusion très grand qui existe entre les deux occlusives bilabiales /p/ et /b/ est-il dû au fait que ces deux sons ne se distinguent que par un seul trait ou doit-on l'attribuer à la nature même du trait?
Miller et Nicely avaient en effet trouvé que les confusions entre consonnes occlusives sont plus fréquentes entre celles qui diffèrent par le lieu d'articulation qu' entre celles qui diffèrent par le voisement.Nos résultats tendraient à laisser penser le contraire.
Au-delà des différences inhérentes aux choix méthodologiques respectifs, cette divergence pose également le problème de la valeur perceptive des oppositions acoustico-phonétique utilisées dans les langues naturelles. Le contraste phonétique entre consonnes sourdes et consonnes voisées dû à la vibration ou à l'absence de vibration des cordes vocales est l'un des plus faciles à produire pour un être humain. Il n'a cependant ni la même valeur phonologique, ni la même réalisation acoustique dans les différentes langues naturelles parlées et perçues par les êtres humains.
La divergence des résultats que nous avons obtenus, avec ceux de Miller et Nicely peut également s'expliquer par le fait que nous avons travaillé sur des langues différentes.
Enfin cette divergence repose à sa manière la question de la validité des concepts de la phonologie-et dans le cas présent du trait-pour les recherches psycholinguistiques tentant de dégager les processus actifs de perception et de compréhension du langage chez l'être humain. Ces résultats laissent penser que, dans une langue donnée toutes les oppositions acoustico-phonétiques entre les sons -ici, le voisement en français- ne sont pas traités perceptivement de manière identique par les locuteurs de cette langue,même si elles ont toutes la même valeur distinctive.Cette question mérite cependant d'être approfondie par des recherches expérimentales systématiques.

CONCLUSION

L'existence et la nature des invariants acoustiques de la parole reste un problème ouvert et insoluble à l'heure

actuelle. Les confusions ne peuvent pas être abordées uniquement par l'étude du signal de parole.Quant aux quelques recherches proposées à ce jour en perception,elles ont le plus souvent recours au jugement métalinguistique des sujets.
Nous proposons ici une tâche et une procédure expérimentales qui permettent d'étudier ce phénomène en temps réel. Cette recherche constitue un premier pas dans un domaine qui mérite d'être approfondi, notamment par une étude systématique de la valeur perceptive respective et comparée des oppositions distinctives du système phonologique du français.Ces résultats affineraient sans doute les connaissances que nous avons des confusions perceptives qui sont le plus souvent considérées comme équivalentes quel que soit le contexte dans lequel elles se réalisent, et alors qu'elles ne se produisent pas de manière aléatoire. Ceci permettrait également d'éclairer les modèles de perception de la parole.

BIBLIOGRAPHIE

Dell G.S and Newman J.E.(1980). Detecting phonemes in fluent Speech.Journal of verbal learning and verbal behavior ,19,608623.

Foss D.J. and Swinney D.A.(1973). On the psychological reality of the phoneme:Perception,identification and consciousness.Journal of verbal learning and verbal behavior,12, 246-257.

Miller G.A. et Nicely P.E. (1955).analyse de confusions perceptives entre consonnes anglaises.Traduction et adaptation de l'anglais in J.Mehler et G.Noizet(eds) .Textes pour une Psycholinguistique. Mouton,Paris La Haye,1974.

Newman J.E. and Dell G.S. (1978)The phonological nature of phoneme monitoring:a critique of some ambiguity studies.Journal of verbal learning and behavior ,17,359-374.

19es J.E.P. - Bruxelles, 19 au 22 mai 1992.

MODÈLES D'INTÉGRATION AUDITION-VISION DANS LA PERCEPTION DES VOYELLES

Jordi Robert-Ribes, Pierre Escudier, Jean-Luc Schwartz

Institut de la Communication Parlée
URA CNRS n°368; INPG - Université STENDHAL
INPG, 46 Av. Félix Viallet 38031 Grenoble Cedex FRANCE

Résumé

La perception de la parole est à la fois auditive et visuelle, il faut donc résoudre le problème de l'intégration perceptive de ces deux modalités. Plusieurs modèles d'intégration ont été proposés dans la littérature. Ces modèles sont ici expliqués, et il apparait que seuls peuvent être acceptés ceux qui s'appuient à un niveau précédant le processus de décision (identification) sur une *représentation commune* des stimulations de chacune des modalités. Nous avons alors testé, sur la perception audiovisuelle de voyelles monolocuteur, deux modèles, implémentés à base de réseaux neuromimétiques. Le premier modèle associe à l'image un son, puis intègre les deux informations dans un espace de représentations spectrales. Le deuxième modèle élabore à partir du son et de l'image une représentation de la forme du conduit vocal, à partir de laquelle se fait l'identification. Les modèles donnent en l'état actuel des résultats encourageants.

1. INTRODUCTION

L'enjeu central de la perception de parole est la récupération d'un code phonologique à partir d'un ensemble de signaux physiques, eux-mêmes produits de manoeuvres articulatoires du locuteur destinées à transmettre ce code. Ces signaux physiques sont, bien sûr, du son — des signaux acoustiques — mais aussi de l'image, celle du "visage parlant". Il nous faut donc comprendre comment s'effectue cette "combinaison" d'informations de diverses modalités. Le présent travail concerne la perception audiovisuelle des voyelles du français. Différents modèles seront exposés dans la Section 2. La Section 3 proposera une stratégie de recherche pour cette étude, aboutissant à la présentation des modèles retenus et des tests mis en oeuvre pour les évaluer. La Section 4 décrira les résultats obtenus avec chacun des modèles étudiés. Pour finir nous donnerons dans la Section 5 quelques réflexions de conclusion sur les différents modèles.

2. MODÈLES D'INTÉGRATION AUDITION-VISION (Summerfield, 1987)

2.1 Traits phonétiques

Ce modèle propose que les représentations auditive et visuelle sont d'abord catégorisées, l'intégration se réalisant ensuite en prenant des traits d'après la catégorisation de la représentation visuelle, et d'autres traits d'après la catégorisation de la représentation auditive. Dans la proposition initiale d'intégration par traits, on suppose que la modalité visuelle fournit exclusivement les traits phonétiques de *lieu d'articulation* et la modalité auditive ceux de *mode d'articulation* (hypothèse VPAM pour "Vision Place, Audition Manner").

C'est le seul des cinq modèles proposés par Summerfield où les informations sont intégrées *après* catégorisation. Dans les autres cas, on intègre des fonctions analogiques et continues. Cependant, les données expérimentales (Massaro, 1987 ; Braida, 1991) montrent qu'un tel mécanisme d'intégration de valeurs logiques n'est pas tenable. La logique floue permet de résoudre ce problème (Massaro, 1987).

2.2 Fonction de transfert du conduit vocal

Dans cette hypothèse, on suppose que le système visuel "calcule" à chaque instant une fonction de transfert du conduit vocal ; de son coté, le système auditif calcule, aussi, une fonction de transfert, et donne de plus de l'information sur le type de source (voisé/non-voisé). Une moyenne des deux fonctions de transfert est alors estimée, et avec l'information sur la source on catégorise en phonèmes. On fait donc l'hypothèse qu'il existe une "modalité dominante", l'audition en ce qui concerne la perception de la parole, avec un recodage de la modalité "faible" (la vision) dans un espace de représentation de la modalité dominante (représentation spectrale).

2.3 Spectres acoustiques et formes visibles

Ce modèle propose que les stimuli auditifs et visuels sont représentés dans un espace bimodal. Les points de cet espace bimodal sont convertis en phonèmes (donc identifiés) par comparaison avec des prototypes dans un dictionnaire bi-vectoriel (chaque stimulus est

représenté par un vecteur dans l'espace son x image). On travaille donc ici directement sur un espace de "sensations", sans faire appel à un niveau intermédiaire de représentation commun aux modalités auditive et visuelle

2.4 Configurations du conduit vocal

Dans ce modèle, chaque modalité fournit des caractéristiques de la fonction d'aire du conduit vocal. De la fonction d'aire on passe ensuite aux phonèmes par catégorisation. On effectue donc sur chacune des deux modalités un transcodage vers un espace de représentation "du 3ème type", apparenté non à l'une des deux entrées perceptives, mais à un espace de *commandes motrices* : l'intégration se fait par référence aux relations perception / action (Hatwell, 1986, p.53-78).

2.5 Dynamique articulatoire

Les modèles 2.2 et 2.4 prennent en compte la configuration spatiale des articulateurs, mais non leur dynamique. Au contraire, ce dernier modèle propose de tenir compte du fait que l'interaction entre vision et audition n'est pas constante, mais dépend de la dynamique des deux entrées. Ce sont les principes de contrôle de cette dynamique qui doivent être récupérés par l'audition et la vision. Rien n'est dit par Summerfield sur la nature de ces paramètres dynamiques (mais voir Schwartz et al., 1992). Par la suite, nous ne considérerons plus ce modèle qui reste assez vague et ne présente pas de différence avec le modèle 4 (Section 2.4) dans le cas de voyelles stationnaires.

3. STRATÉGIE D'ÉVALUATION DES MODÈLES SUR DES STIMULI VOCALIQUES

3.1 Principe général d'implémentation par réseaux neuromimétiques

Tous les modèles que nous avons présentés passent par une étape de recodification ou d'association. Ce type de tâche convient bien aux réseaux neuromimétiques avec rétropropagation du gradient d'erreur (Rumelhart et al., 1986). Chaque modèle, une fois réglé par apprentissage, devra être capable de réaliser, à base d'associations neuromimétiques et de mécanismes d'intégration que nous préciserons ultérieurement, le passage d'une entrée auditive ou d'une entrée visuelle ou d'un couple entrée auditive – entrée visuelle à un ensemble de probabilités de réponses associées respectivement à chacune des catégories vocaliques du français (nous avons utilisé 10 catégories en sortie : [a, ɛ, e, i, œ, ø, y, ɔ, o, u] et consideré [a] comme une voyelle d'avant).

3.2. Faits expérimentaux à reproduire

Nous nous sommes appuyés sur deux paradigmes principaux : la perception dans le bruit et la perception de stimuli conflictuels (Robert-Ribes et al., 1992). Nous ne présenterons ici que les résultats de perception de voyelles dans le bruit.

Plusieurs études ont montré que même pour des sujets non-malentendants et non entraînés à la lecture labiale, les informations visuelles peuvent augmenter significativement les taux d'intelligibilité de la parole, notamment pour des signaux bruités. Ainsi, Mohamadi & Benoit (1991) ont montré dans notre laboratoire que, pour un rapport signal sur bruit de -18 dB, le pourcentage de reconnaissance correcte de logatomes ViCjViCjVi avec Vi l'une des 3 voyelles [i, a, y] et Cj l'une des 6 consonnes [b, v, z, ʒ, ʀ, l] passe de 10 %, en présentation auditive seule, à 80 % en présentation audiovisuelle. Ces données montrent le poids considérable de l'entrée visuelle en cas de stimuli bruités. Ceci ne peut être obtenu que par un système d'intégration qui, soit, donne une "prime" systématique aux entrées non ambigües par rapport aux entrées ambigües (voir le Fuzzy-Logical Model of Perception de Massaro, 1987), soit est muni d'un système de *réglage de la pondération audition-vision en fonction du niveau de bruit*, principe proposé par Yuhas et al. (1989) pour le modèle d'intégration dans un espace de spectres (Section 2.2) et facilement transposable au modèle d'intégration dans un espace de gestes (Section 2.4). D'autre part, les données de Summerfield et McGrath (1984) sur la perception de stimuli auditifs et visuels conflictuels suggèrent l'existence d'un *détecteur de conflits* : les modèles doivent être capables de détecter un *écart* entre stimulus visuel et auditif, donc disposer d'un *espace de représentation commun*, et d'un *système de réglage de la pondération audition-vision en fonction du niveau de conflit.*

Ces deux indications semblent incompatibles avec le modèle d'intégration directe (Section 2.3), ne disposant pas d'un espace de représentation commun. Nous décrirons ici les résultats obtenus avec deux modèles qui utilisent chacun un processus en trois étapes : projection des modalités d'entrée vers un espace de représentation commun de nature continue (spatio-temporelle), intégration des deux représentations correspondantes (provenant respectivement de la stimulation auditive et visuelle) puis décodage à partir de la représentation intégrée. Le premier de ces modèles est celui de l'intégration dans un espace de représentations auditives (spectres). Il a déjà été testé avec succès (Yuhas et al., 1989 ; Escudier et al., 1990), mais il n'a jamais été utilisé pour la totalité des voyelles du français (toutes les voyelles arrondies comprises). Le deuxième est le modèle d'intégration dans un espace de représentations articulatoires, étude préliminaire dans le domaine de l'"inversion" (Poggio, 1984 ; Schwartz et al., 1992). On trouvera dans Robert-Ribes (1991) une implémentation du modèle d'intégration dans un espace de traits (Section 2.1) selon les principes de logique floue (Massaro, 1987).

3.3 Méthodologie

3.3.1. Corpus

Le corpus de voyelles est formé de 10 réalisations des 10 voyelles [a, ɛ, e, i, œ, ø, y, ɔ, o, u]. Ces 100 réalisations ont été prononcées par un même locuteur dans une seule scéance d'enregistrement. L'ordre des voyelles était aléatoire. La phrase porteuse était: "C'est pazVze" (où V indique la position de la voyelle). Les signaux ont été numérisés à 10 kHz,

puis, dans chaque logatome, le moment précis (± 10ms) où l'on a considéré la voyelle atteinte a été défini par l'instant où le deuxième formant atteint sa valeur extrêmale à l'intérieur de la trajectoire entre les deux consonnes [z]. Des spectres de 256 points entre 0 et 5 kHz ont alors été calculés à l'instant choisi, par la méthode du cepstre. Puis on a recalculé des spectres sur 32 points, en moyennant arithmétiquement (en dB) des groupes de 8 points des spectres de 256 points. Les spectres ainsi obtenus ont été normalisés pour avoir des valeurs entre 0 et 1, la valeur de 0 dB correspondant à 1 et la valeur de -35 dB à 0. Enfin, un processus de normalisation additive a été effectué, portant la valeur maximale de chaque spectre à 1, de façon à avoir tous les spectres avec la même valeur maximale. Les signaux acoustiques ont été bruités (bruit blanc) avec différents niveaux de rapport signal sur bruit. L'estimation du niveau de bruit et la superposition signal / bruit ont été faites directement dans le domaine spectral.

D'autre part, le visage du locuteur a été filmé et traité par le poste "visage-parole" défini par Lallouache (1990). Pour chaque logatome, l'image correspondant (à une précision de 20 ms) à la fenêtre de calcul du spectre acoustique a été numérisée, puis ramenée, après moyennage, à 120 points avec des niveaux de gris normalisés entre 0 et 1 (0 correspondant au niveau de gris 0 et 1 au niveau de gris 256). On dispose donc, finalement, de 100 paires (spectres 32 points, image 120 points) qui ont fourni la base de nos modèles d'association.

3.3.2 Déroulement du test

Apprentissage

Des dix réalisations disponibles de chaque voyelle, cinq, choisies arbitrairement, ont été utilisées pour faire l'apprentissage (réglage) des réseaux et les cinq autres ont servi à tester les modèles. Les réseaux intervenant dans les différents modèles ont été réglés de façon à s'approcher le mieux possible d'associations idéales que nous définirons par la suite pour chaque modèle. Le choix de la structure de chaque réseau, ainsi que le critère de fin de chaque apprentissage était en général un critère de minimum d'erreur quadratique sur les associations test (limite du "surapprentissage", voir Robert-Ribes (1991) pour plus de détail). Les réseaux comportant comme entrée un spectre ont été réglés avec un corpus *comprenant à la fois des stimuli non bruités et des stimuli bruités.* Ainsi pour chaque voyelle, les cinq réalisations était une fois non bruitées et une autre fois bruitées avec chacune un niveau de bruit différent (18dB, 12dB, 6dB, 0dB, -6dB de RSB). Pour le réglage de ces réseaux un corpus de test du même type a été aussi utilisé (pour les réalisations bruitées, RSB de 18dB, 9dB, 0dB, -9dB et -15dB)

Perception de stimuli acoustiques bruités

Une fois l'apprentissage effectué, chaque réseau a été testé sur des stimuli à des niveaux de bruit s'échellonnant de -42 à 30dB par pas de 3dB. Pour chaque stimulus, seule l'entrée auditive était d'abord utilisée, puis l'entrée visuelle correspondant exactement au signal acoustique dans le corpus enregistré était présentée conjointement au spectre. L'évaluation comparée des performances d'identification dans le bruit dans les deux conditions – avec ou sans entrée visuelle conjointe – pouvait alors être effectuée en sortie de chaque modèle : nous avons défini pour chaque test (1 modèle x 1 condition) un score d'identification globale en associant une note 1 à chaque cas où la sortie la plus excitée correspondait à la catégorie de l'entrée et 0 dans le cas contraire. Nous évaluerons pour les deux modèles l'apport de la vision en fonction du rapport signal sur bruit.

4. RÉSULTATS

4.1 Intégration dans un espace de représentations spectrales (Modèle 1)

4.1.1 Implémentation

Le modèle, rappelons-le, comporte deux voies indépendantes (une pour la vision et une pour l'audition). Chaque voie fournit un spectre (directement pour la voie auditive, par association pour la voie visuelle). Les deux spectres sont intégrés, fournissant un spectre final qui doit être identifié dans l'une des dix catégories vocaliques de notre étude.

Association image-spectre

L'estimation d'un spectre à partir du signal visuel (spectre estimé visuellement) peut être faite en utilisant des réseaux neuromimétiques. Nous avons cependant rencontré un problème particulier en ce qui concerne les voyelles du français : c'est le problème de la non-biunivocité entre visèmes et phonèmes, due à l'existence de voyelles d'avant arrondies et voyelles d'arrière arrondies. Ces voyelles forment des paires qui ont le même visème (même forme des lèvres), et des fonctions de transfert très différentes. Ce point n'a pas été abordé par Summerfield, puisque les voyelles de l'anglais ne présentent pas la distinction phonologique arrondi – non arrondi.

Ce problème a été résolu en associant deux spectres à chaque image. Pour les visèmes arrondis, ce sont respectivement le spectre correspondant à la voyelle d'avant arrondie et celui correspondant à la voyelle d'arrière arrondie. Pour les visèmes non-arrondis, les deux spectres correspondent à la même voyelle avant non-arrondie. Nous avons donc défini deux réseaux associateurs image-spectre, respectivement nommés "réseau antérieur" et "réseau périphérique". Le réseau antérieur, fournissant un "spectre antérieur", associe à des visèmes arrondis des spectres de voyelles d'avant arrondies et à des visèmes non arrondis des spectres de voyelles d'avant non arrondies, et le réseau périphérique, fournissant un "spectre périphérique", associe à des visèmes arrondis des spectres de voyelles d'arrière et à des visèmes non arrondis des spectres de voyelles d'avant non arrondies. A l'apprentissage, le réseau antérieur a été reglé avec les sept voyelles d'avant [a, ɛ, e, i, œ, ø, y], tandis que le réseau périphérique a été reglé avec les sept voyelles périphériques [a, ɛ, e, i, ɔ, o, u]. Dans chaque réseau, une seule couche cachée a été utilisée, et des essais

préliminaires ont conduit à l'utilisation de 20 neurones cachés.
En phase de test, à une image inconnue sont donc associés deux spectres et il faut en choisir un des deux pour le pondérer avec le spectre acoustique. Le choix est fait à partir de la distance auditive proposée par Bladon (1981). Ainsi, le spectre estimé visuellement qui minimise cette distance calculée avec le spectre acoustique est pondéré avec ce dernier. En d'autres termes, dans le cas de visèmes arrondis, la décision "spectre de voyelle d'avant" (fourni par le réseau antérieur) vs "spectre de voyelle d'arrière" (fourni par le réseau périphérique) est faite sur la base d'une comparaison avec le spectre acoustique correspondant, qui contient une partie haute fréquence beaucoup plus importante dans le cas des voyelles d'avant (F2 plus élevé).

Intégration

La pondération entre spectre acoustique et spectre visuel estimé est faite point par point selon la formule :

$$S_{n,\text{pondéré}} = \alpha \cdot S_{n,\text{acoustique}} + (1-\alpha) \cdot S_{n,\text{visuel}}$$
$$n \in \{1,..,32\},\ \alpha \in [0,1]$$

La pondération dépend d'un facteur α fonction du rapport signal sur bruit. Nous avons dans l'immédiat repris le type de variations de α avec S/N proposé par Yuhas et al. (1989), en l'adaptant à nos données. Nous avons ainsi choisi :

$$\begin{cases} \alpha = 1.0 & S/N > -6dB \\ \alpha = 0.03 \cdot RSB + 1.18 & -36 < S/N < -6dB \\ \alpha = 0.1 & S/N < -36dB \end{cases}$$

Nous envisageons dans une étape ultérieure d'optimiser α par apprentissage.

Identification

Une fois obtenu le spectre résultant de l'intégration, on lui associe des niveaux d'excitation correspondant à chacune des dix catégories vocaliques par un autre réseau neuromimétique d'identification. Ce second réseau a été réglé par l'apprentissage du passage entre spectres acoustiques et réponses parfaites (1 pour la classe considéré, 0 pour les autres classes, avec cinq associations à apprendre pour chacune des dix classes). La structure du modèle est représentée dans la Fig. 1a.

4.1.2 Résultats

Selon le protocole défini dans la Section 3.4.2, on obtient (Fig. 1b) les variations de scores de reconnaissance, sans et avec apport de la voie visuelle, en fonction du niveau de bruit.
Nous les commenterons plus en détail dans la discussion générale (Section 5). Notons que, si seuls les signaux visuels sont présentés en entrée, le réseau ne pourra pas distinguer, pour les voyelles arrondies, entre voyelle d'avant et voyelle périphérique. Un test réalisé séparément pour les voyelles d'avant et les voyelles périphériques montre que l'on peut obtenir dans chaque ensemble (sept catégories par ensemble) un score de 84 % d'identification visuelle à partir des seules images des lèvres. Ces scores sont relativement élevés. Néanmoins, sur ce problème simple (voyelles bien prononcées à contexte fixe, un seul locuteur, images de bonne qualité, choix parmi sept catégories), on pourrait s'attendre à des scores encore supérieurs. Il faut noter toutefois que la procédure d'identification n'est pas optimale, puisqu'il ne s'agit pas d'une procédure de reconnaissance des formes appliquée directement sur l'image, mais d'une reconnaissance en deux temps : on cumule donc les imprécisions du réseau associateur image-spectre et du réseau identificateur spectre-catégorie vocalique.

4.2 Intégration dans un espace de représentations articulatoires (Modèle 2)

Le problème principal qui se pose pour la conception d'un tel modèle est celui du choix d'un *espace de représentation du conduit vocal*. Finalement (voir Robert-Ribes, 1991) nous avons opté pour un espace de contrôle "classique" décrivant la configuration articulatoire par trois paramètres *continus* : arrondissement, degré de hauteur, degré de recul sur un axe avant-arrière ("postériorité" dans la suite du texte).

4.2.1 Implémentation

Ce modèle présente trois étapes bien définies : obtention des paramètres représentant le conduit vocal dans l'espace défini précédemment, intégration de ces paramètres pour les deux modalités, et classification.

Obtention des paramètres

A partir de l'image, deux paramètres ont été obtenus, un pour l'arrondissement et un pour la hauteur. Le spectre donne trois paramètres : arrondissement, hauteur et postériorité. Ces paramètres sont des valeurs continues décrivant une caractéristique quantitative de la configuration articulatoire. Ainsi, le paramètre d'arrondissement varie entre 0 pour un degré d'arrondissement minimal et 1 pour un degré d'arrondissement maximal, le paramètre de postériorité entre 0 indiquant "le plus avant" et 1 pour "le plus arrière", cependant que le paramètre de hauteur varie continuement entre 0 et 1 en passant des voyelles basses aux voyelles mi-basses, mi-hautes puis hautes. Par exemple, à [a] correspondent à l'apprentissage 0 pour le paramètre d'arrondissement, 0 pour la hauteur et 0 pour la postériorité (notre choix, rappelons-le, étant de faire de [a] une voyelle d'avant) et [u] est défini par les valeurs 1 pour l'arrondissement, 1 pour la hauteur et 1 pour la postériorité. Une nouvelle fois, toutes les évaluations de paramètres à partir d'entrées auditives ou visuelles ont été réalisées par réseaux neuromimétique.

Intégration des paramètres

Le paramètre de postériorité provient de la voie auditive seule, tandis que les paramètres de hauteur et d'arrondisement sont obtenus par somme pondérée des paramètres estimés à l'étape précédente. La pondération est de la même forme que pour le modèle d'intégration de représentations spectrales, c'est-à-dire :

$$P_{\text{pondéré}} = \alpha \cdot P_{\text{acoustique}} + (1-\alpha) \cdot P_{\text{visuel}}$$

P étant le paramètre correspondant (arrondissement ou hauteur), et α le coefficient de pondération.

Identification

Nous avons défini pour chaque paramètre articulatoire des zones correspondant à la réalisation des traits: arrondi, non-arrondi, bas, mi-bas, mi-haut, haut,

antérieur, posterieur. Les réalisations donnant comme sortie finale pondérée des valeurs des paramètres comprises entre les limites correspondantes à sa classe vocalique ont été considérés comme reconnues correctement. On en arrive au schéma de la Fig. 2a.

4.2.2 Résultats

Perception dans le bruit

Nous avons utilisé une fonction de pondération des différents paramètres auditifs et visuels selon la formule :

$$\begin{cases} \alpha = 0.8 & S/N > 21dB \\ \alpha = 0.022 \cdot RSB + 0.33 & -15 < S/N < 21dB \\ \alpha = 0.0 & S/N < -15dB \end{cases}$$

Par la suite, nous envisageons, là encore, d'optimiser α par apprentissage.

Les résultats sont portés dans la Fig. 2b. Notons que, à partir de l'information visuelle seule, l'identification des voyelles ne peut être obtenue, puisque l'information sur la position (avant-arrière) n'est pas disponible, mais que si l'on donne l'information avant-arrière, le pourcentage d'identification correcte est de 94%.

5. DISCUSSION ET CONCLUSION

Les deux modèles que nous avons testés fournissent d'ores et déjà des résultats prometteurs. Nous avons proposé dans chaque cas une solution pour résoudre le problème spécifique de l'existence en français de visèmes identiques pour des phonèmes différents, et nous avons finalement défini deux structures répondant bien à nos exigences de départ, c'est-à-dire capables d'associer une catégorie vocalique aussi bien à un spectre qu'à une image ou à une paire spectre-image, en passant par un espace de représentation commun aux modalités auditive et visuelle, et de nature continue (précédant l'identification). Enfin, nous avons montré dans l'un et l'autre cas des bonnes capacités d'identification de voyelles dans le bruit. Il reste cependant encore un important travail d'optimisation de nos modèles portant principalement sur trois points : (1) architecture des réseaux neuromimétiques mis en oeuvre et test éventuel d'autres méthodes d'association ou d'identification, (2) optimisation du mécanisme d'intégration lui-même, en faisant rentrer par exemple la variable α à l'intérieur du processus d'apprentissage, et (3) recherche du meilleur domaine d'apprentissage dans le bruit.

La comparaison des Fig. 1b et 2b montre peu de différences entre les deux modèles, avec un léger avantage au modèle 1 (Fig. 1b), avantage qui reste à confirmer après mise en oeuvre des travaux que nous venons d'indiquer. Il nous semble cependant que le modèle 2 est plus prometeur pour l'avenir pour 4 raisons.

(1) Il permet un *meilleur traitement de la modalité visuelle*, à laquelle on associe directement une représentation proche de la description en traits phonétiques au lieu de passer par un équivalent spectral. Ceci se traduit d'ailleurs déjà sur nos résultats présents : on obtient pour la modalité visuelle seule un meilleur score d'identification dans le modèle 2 que dans le modèle 1, pour lequel les imprécisions des deux réseaux d'association image-spectre et d'identification de spectres se cumulent.

(2) La nature *paramétrique* de la représentation articulatoire que nous avons choisie permet une exploitation plus aisée de la complémentarité *audition-vision*, c'est-à-dire du fait que certains traits sont plus facilement identifiables auditivement, et d'autres visuellement (Summerfield, 1987).

(3) Cette nature paramétrique évite, dans le cas d'intégration de stimuli auditifs et visuels conflictuels, la *génération de "monstres"*, ce qui n'est pas le cas du modèle 1 pour lequel l'intégration d'un spectre "auditif" et d'un spectre "estimé visuellement" de forme assez différente conduit à la fabrication de monstres spectraux ininterprétables (Robert-Ribes et al., 1982).

(4) Il permet d'aborder le problème de la *mise en correspondance de percepts et de représentations articulatoires*, problème clé pour l'étude des mécanismes d'apprentissage de gestes à partir de percepts auditifs et visuels.

Il reste cependant que notre travail de mise en oeuvre et de comparaison de modèles passe à l'avenir par la confrontation à deux types de probleme, celui de la variabilité interlocuteur - et donc de la *normalisation* dans les représentations auditives, visuelles et articulatoires - et celui du *traitement de stimuli dynamiques* - et donc du choix de représentations adéquates dans ces différents espaces.

Bibliographie

Bladon, R.A.W., & Lindblom, B. (1981) Modelling the judgement of vowel quality differences. *J. Acoust. Soc. Am., 69*, 1414-1422.

Braida, L. (1991) Crossmodal integration in the identification of consonant segments. A paraître dans Q. Summerfield (Ed.) *Hearing and Speech, A special Issue of the Quaterly Journal of Experimental Psychology*. Lawrence Erlbaum Associates Ltd.

Hatwell, Y. (1986) *Toucher l'espace. La main et la perception tactile de l'espace.* Presses Universitaires de Lille.

Lallouache, T. (1990) Un poste "Visage-Parole". Acquisition et traitement de contours labiaux. *18èmes JEP*, pp 282-286.

Massaro, D.W. (1987) *Speech perception by ear and eye: A paradigm for psychological inquiry.* Hillsdale, NJ: Erlbaum, London.

Mohamadi, T., & Benoït, C. (1991) Apport de la vision du locuteur à l'intelligibilité de la parole bruitée. Proposé au *Bulletin de la Communication Parlée.*

Robert-Ribes, J. (1991) *Intégration audition-vision par réseau de neurones : une étude comparative des modèles d'intégration appliqués à la perception des voyelles.* Rapport de stage, DEA Signal-Image-Parole, INP Grenoble.

Robert-Ribes, J., Escudier, P., Schwartz, J.L., (1992) Modèles d'intégration audition-vision dans la perception des voyelles: une étude neuromimétique. *5ème Colloque de l'ARC*, Nancy, mars 1992.

Rumelhart, D.E., Hinton G.E., Williams R.J. (1986) Learning Internal Representations by Error Propagation. In D.E. Rumelhart & J.L. McClelland (Eds.), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Vol 1: Foundations.* MIT Press.

Schwartz, J.L., Arrouas, Y., Beautemps, D., & Escudier, P. (1992) Auditory analysis of speech gestures. In B. Schouten (ed.) *The Processing of Speech : from the Auditory Periphery to Word Recognition.* Berlin : Mouton-De Gruyter (à paraître).

Summerfield, Q. (1987) Some preliminaries to a comprehensive account of audio-visual speech perception. In B. Dodd & R. Campbell (Eds.), *Hearing by eye : the psychology of lipreading* (pp. 3-51). Lawrence Erlbaum Associates, London.

Summerfield, Q., & Mc Grath, M. (1984) Detection and resolution of audio-visual incompatibility in the perception of vowels. *Quarterly Journal of Experimental Psychology : Human Experimental Psychology, 36(1-A)*, 51–74.

Yuhas, B.P., Goldstein, M.H., Jr., & Sejnowski, T.J. (1989) Integration of acoustic and visual speech signals using neural networks. *IEEE Communications Magazine*, Nov. 1989, 65-71.

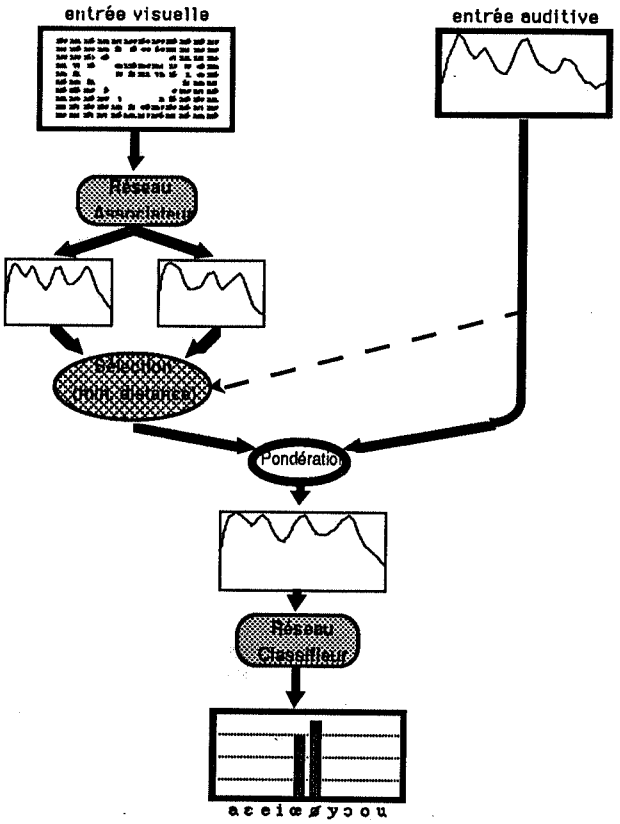


Figure 1a– Schéma général du modèle d'intégration de représentations spectrales

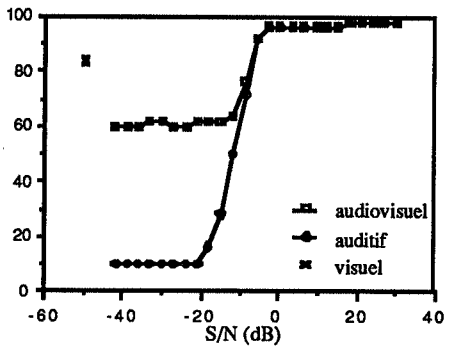


Figure 1b – Résultats modèle d'intégration de représentations spectrales (scores d'identification visuelle, auditive et audiovisuelle)

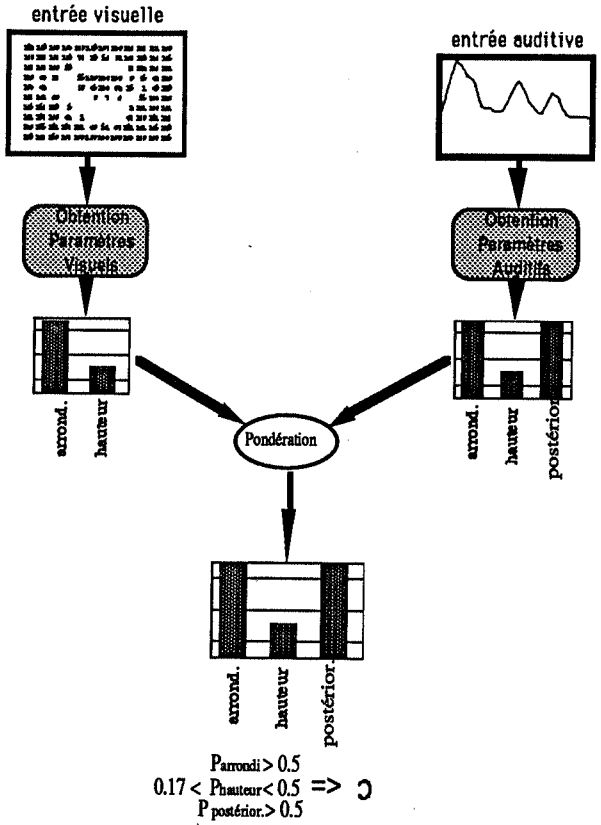


Figure 2a – Schéma général du modèle d'intégration de représentations articulatoires

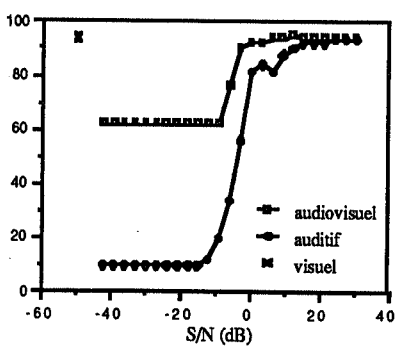


Figure 2b – Résultats modèle d'intégration de représentations articulatoires (scores d'identification visuelle, auditive et audiovisuelle)

UNE PRÉDICTION DE L' « AUDIBILITÉ » DES GESTES DE LA PAROLE À PARTIR D'UNE MODÉLISATION ARTICULATOIRE

Louis-Jean BOË, Pascal PERRIER & Andrew MORRIS

INSTITUT DE LA COMMUNICATION PARLÉE
URA CNRS n° 368 INPG/ENSERG Université Stendhal BP 25 38040 Grenoble Cedex 9 France

Résumé

La non-linéarité entre le mouvement des articulateurs et les paramètres acoustiques laisse prévoir les conséquences spectrales différenciées que peuvent avoir certains gestes articulatoires. À partir d'un modèle de production à 7 degrés de liberté (MAEDA, 1989) ont été générées les *macro-variations acoustiques* correspondant à 10 prototypes de voyelles cardinales. L'analyse systématique de ces simulations permet de mettre en évidence 3 types de non-linéarité : *très faibles dépendances* pour la protrusion labiale et la position du larynx, *saturations* dues pour l'essentiel à la séparation labiale et enfin *non-monotonies* de type parabolique avec les mouvements du dos de la langue et de l'apex. Ces phénomènes sont discutés du point de vue de la coordination des gestes : les gestes *peu audibles* autorisent l'anticipation et une certaine latitude de production, et peuvent correspondre à des manœuvres synergiques facilitantes ; inversement les mouvements *très audibles* jouent le rôle déterminant dans la réalisation des cibles acoustiques.

1. INTRODUCTION

La relation entre déplacements articulatoires et caractéristiques acoustiques n'est pas linéaire : certains gestes de production sont plus *audibles* que d'autres. La *Théorie Quantique* de STEVENS (1972, 1989) – largement rediscutée récemment (OHALA, 1989) – a fait de ce phénomène une des propriétés centrale du processus de communication linguistique. Le code phonétique pourrait ainsi se structurer par rapport aux zones de forte non-linéarité qui assurent une bonne distinctivité acoustico-perceptive. D'autre part, du point de vue de la coordination des gestes, cette possibilité pourrait être exploitée dans les phénomènes d'anticipation (ABRY & SCHWARTZ, 1988/89). Enfin, il est possible que certains gestes non audibles autorisent des manœuvres articulatoires facilitantes permettant d'assurer, de manière synergique, une meilleure précision de la production acoustique. Inversement, certains gestes très audibles ont une importance cruciale pour le contrôle de la cible acoustique.

Pour tenter d'avancer dans la compréhension de ces phénomènes, un modèle intégrant les contraintes articulatoires, avec comme paramètres d'entrée les 4 principaux articulateurs, nous a semblé bien adapté. Les 7 degrés de liberté du modèle de MAEDA (1989) – les lèvres (séparation et protrusion), la mâchoire, la langue (corps, dos, apex), le larynx – extraits à partir des données radiographiques et labiographiques, permettent de générer les variations géométriques du conduit vocal. Chacun des paramètres apporte un pourcentage d'explication de la variance de l'ensemble des formes géométriques qui ont servi à l'analyse (1 000 au total, correspondant à dix phrases prononcées à débit normal) (BOTHOREL & al., 1986). Les composantes corps, dos, apex et mâchoire expliquent 88% de la variance de la forme géométrique de la langue (respectivement 43%, 23%, 7% et 15%). La mâchoire, la hauteur et la protrusion labiale expliquent 100% de la variance de la géométrie du pavillon labial (de face et de profil) ; la hauteur du larynx et la mâchoire contrôlent la position verticale du larynx. Le domaine de variation de chaque articulateur est normalisé par rapport à l'écart-type (σ) calculé à partir des données radiographiques. Le modèle de MAEDA a été implanté sous forme de *document interactif* permettant à l'utilisateur, à partir de la manipulation des commandes (cf. par exemple SLANEY, 1990), d'obtenir une simulation présentant, graphiquement et numériquement, les données articulatoires et acoustiques (spectrales et temporelles) associées (figure 1). Pour 10 prototypes de voyelles cardinales correspondant à [i, e ɛ, a, y, ø, œ, u, o, ɔ] (VALLÉE & BOË, 1992) nous avons généré les *macro-variations* acoustiques, paramètre par paramètre (gamme de variation de $\pm 3\sigma$ et par pas de 0.25σ,

excepté pour le larynx ± 1σ et pas de 0.1σ). Pour rester dans un domaine de production vocalique, nous avons arrêté l'exploration avant l'apparition d'une occlusion, comme par exemple dans le cas d'un relèvement trop important du dos ou de l'apex, ou de la fermeture des lèvres. Nous avons ainsi obtenu un dictionnaire de 1 421 items contenant, pour chaque réalisation (en moyenne 20 par voyelles), les valeurs des 7 paramètres et celles des quatre premiers formants (Figure 2). Nous avons examiné systématiquement, pour chaque voyelle, et paramètre par paramètre, les relations qui lient espace articulatoire et espace acoustique. À partir de celles-ci il est possible d'opérer une lecture de l'influence acoustique des gestes.

2. UNE NON-LINÉARITÉ CERTAINE

L'influence acoustique des gestes est fortement dépendante des relations de non-linéarité, et pour les mettre en évidence globalement, nous avons utilisé une optimisation polynomiale, au sens des moindres carrés, (MORRIS, 1990) de degré 1 :

$$F_j(P) = w_{oj} + \sum_{i=1}^{7} w_{ij} P_i$$

Les erreurs quadratiques moyennes relatives (erreur quadratique / écart-type) sont respectivement de 43%, 18%, 46% et 54% pour les quatre premiers formants. Une relation linéaire entre paramètres de commande et formants, évaluée avec ce dictionnaire vocalique, entraîne donc, au mieux 18% et au pire 54% d'erreurs : il s'agit bien là d'une non-linéarité importante entre entrée et sortie du modèle. En fait cette estimation globale demande à être affinée, paramètre par paramètre, voyelle par voyelle. Le R^2 (carré du coefficient de corrélation linéaire) correspond au rapport entre la variance estimée et la variance observée. Calculé pour les 10 prototypes du français, les 7 paramètres et les 4 formants (280 réalisations, au total) il permet de préciser que dans 40% des cas, le pourcentage de la variance relative expliquée par un modèle linéaire est inférieur à 90%. On observe alors les types de non-linéarité de la figure 3 : indépendance totale (ou plutôt faible dépendance), saturation à la commande et non-monotonie parabolique.

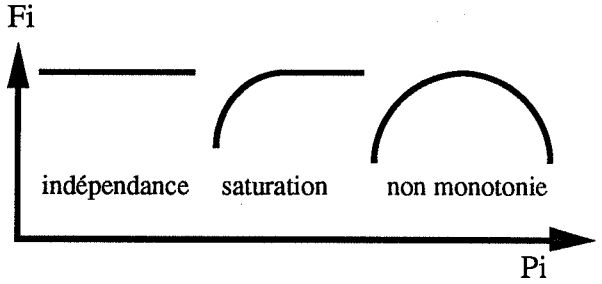


Figure 3 : Type de relations non-linéaires schématisées entre paramètres de commande et formants.

Le tableau 1 présente les relations non-linéaires dans les cas où $R^2 < 0.50$. Typiquement la séparation labiale (*Lip Height*) présente une non-linéarité bien connue de type "saturation" (ABRY & al., 1989 ; BOË & PERRIER, 1991), alors que le Dos (*Drsm*) et l'Apex induisent une variation acoustique non-monotone, invalidant une relation linéaire. On n'obtient pas de véritable indépendance entre commande et formants, mais seulement, comme nous le verrons, de faibles coefficients de variation pour la protrusion et la hauteur du larynx.

	F1	F2	F3	F4
Lip H	i y	i y	i y	i e y ɔ
Jaw		ɔ		
Body	y		u y ɔ	i e ɛ a ø œ u o
Drsm	i y ø œ o	e ɛ a y œ ɔ		
Apex	œ		a	

Tableau 1 : Voyelles présentant des relations non-linéaires entre paramètres et formants, avec $R^2 < 0.50$ (variance prédite < 50%).

3. PRÉDICTION DE L'AUDIBILITÉ

Le coefficient de variation du formant F_j:

$$100\ (\Sigma\ (F_{j\ cible} - F_{i\ j})^2 / n)^{1/2} / F_{j\ cible}$$

calculé pour chaque macro-variation (par paramètre et par voyelle) peut nous permettre d'évaluer les conséquences acoustiques des gestes. Si l'on se réfère aux seuils de perception établis pour les formants des voyelles (FLANAGAN, 1955 ; MERMELSTEIN, 1978) on peut supposer qu'une variation de 15% pour F1 et/ou de 10% pour les autres sera perceptible.

3.1 Gestes peu audibles

Avec les contraintes de seuil appliquées aux 4 premiers formants, on relève les gestes peu audibles suivants :

- Séparation labiale pour [i, e]
- Protrusion labiale pour toutes les voyelles
- Mâchoire pour [o, u]
- Dos pour [i, e, ɛ, a, ø, œ, u, o]
- Apex pour [i, y, œ, u]
- Larynx pour [i, e, ɛ, a, y, ø, œ, u, o, ɔ]

La présence de la séparation labiale dans cette liste n'est pas pour nous surprendre, il est bien connu que les deux voyelles fermées d'avant peuvent avoir des aires labiales très variables (ABRY & BOË, 1986 ; ZERLING, 1990 ; BENOIT & al., 1991). Le dernier modèle de MAEDA (version à 7 paramètres), en introduisant deux degrés de liberté pour les lèvres permet de mettre en évidence un phénomène important : le geste d'étirement-protrusion, signalé dans toutes les descriptions vocaliques, a très peu d'influence acoustique sur les 10 voyelles. Nous faisons l'hypothèse que ce geste, s'il n'est pas audible, est utilisé pour la précision du contrôle de l'aire labiale. Il est en effet

possible de produire des [y, u] lèvres non protruses, mais, dans ce cas, le contrôle de l'aire aux lèvres, s'il est assuré par le seul déplacement de la mâchoire, est beaucoup moins précis sans la proprioception qu'offre la musculature labiale. Autre geste peu audible, celui de la mâchoire pour les 2 voyelles fermées d'arrière : on se trouve dans une forte zone de saturation, déjà soulignée par MAJID (1986). Les non-linéarités introduites par les mouvements de l'apex sont à saturation ou non monotones. La position du larynx n'a que peu d'influence sur l'ensemble des voyelles, bien que certains déplacements soient facilitants (abaissement pour diminuer F2 de [y], phénomène relevé par exemple par RIORDAN, 1977 ; WOOD, 1986) ; il est vrai qu'au niveau articulatoire son influence est très localisée. Par rapport à leurs importances géométriques dans le modèle, ce sont surtout la protrusion labiale et le dos de la langue qui sont les grands perdants acoustiques. On peut prévoir que des gestes d'anticipation peuvent les affecter. Pour les lèvres (LALLOUACHE, 1991), ces phénomènes commencent à être systématiquement étudiés en dynamique (Figure 4, pour des exemples de non-linéarité).

3.2. Gestes audibles

Le tableau 2 présente les coefficients de variations, au dessus des seuils de perception, pour chacune des voyelles et par formants. Plusieurs paramètres se partagent la vedette : incontestablement, toutes voyelles et formants confondus, c'est le corps de la langue qui est en tête. Son déplacement d'avant en arrière a un effet toujours significatif sur F2, mais aussi sur F1 de [i, e, œ, ɔ] et F3 de [i, e, ɛ, a, ø, œ]. Arrivent ensuite, distanciées, la séparation labiale, influente sur F1 de [ɛ, a, y, ø, œ, u, o, ɔ], sur F2 de [u,o], sur F3 de [y]; et la mâchoire pour F1 de [i, ɛ, y, ø, œ], F2 de [ɛ, a, ø, œ, ɔ]. Enfin, le dos est essentiellement opérant sur F3 de [e, y, u].

4. DISCUSSION

Au passage remarquons que les macro-variations sont d'utilité pour alimenter la discussion sur certains points délicats. Par exemple notons les gestes qui facilitent la production du contraste [i] *vs.* [y] (abaissement de F2 et surtout de F3 : MANTAKAS, 1989) : essentiellement diminution de la séparation labiale, mais aussi abaissement du dos et du larynx, léger recul du corps de la langue.

Mais le phénomène important quantifié ici c'est la non-linéarité qui pondère très nettement les relations entre articulateurs et cibles acoustiques.

• Certains gestes révèlent bien leur rôle crucial pour la production : le corps de la langue pour toutes les voyelles, la séparation labiale pour toutes les voyelles sauf [i, e], la mâchoire pour [i, e, ɛ, a, y, ø, œ, ɔ]. Le corps de la langue est en effet directement responsable du lieu d'articulation dont on connaît toute la stabilité et l'importance typologique. La mâchoire intervient directement sur la dimension de la constriction et confirme son rôle, longtemps contesté, d'articulateur difficilement contournable pour la modélisation.

• À l'inverse, la protrusion, la position du larynx et le dos de la langue sont les moins audibles. Ces deux derniers paramètres participaient peu pour rendre compte de la variance géométrique observée par MAEDA, alors que nous aurions *a priori* surestimé l'effet de la protrusion, souvent associée dans les descriptions à l'écartement labial. Ils sont potentiellement utilisables pour des stratégies de coarticulation de protrusion et d'aperture. Pour récupérer ces gestes à partir du signal acoustique, il faudra dépasser le niveau géométrique et intégrer à dans l'inversion des principes de contrôle moteur (minimisation « d'effort », synergie...).

	F1	F2	F3
i	*Body* 95 *Jaw* 25	*Body* 35	*Body* 23
e	*Body* 22 *Apex* 20 *Jaw* 19 *Lip H* 15 *Drsm* 15	*Body* 31	*Body* 13 *Drsm* 10
ɛ	*Jaw* 28 *Lip H* 22 *Body* 21 *Apex* 18	*Body* 27 *Jaw* 14 *Lip H* 10	*Jaw* 17 *Body* 13
a	*Lip H* 36 *Apex* 27 *Body* 24 *Jaw* 16	*Body* 38 *Jaw* 15	*Body* 11
y	*Jaw* 23 *Lip H* 15	*Body* 35	*Lip H* 16 *Drsm* 12
ø	*Lip H* 19 *Jaw* 16 *Body* 16	*Body* 28 *Jaw* 13 *Apex* 10	*Body* 10
œ	*Body* 25 *Lip H* 23 *Jaw* 20	*Body* 33 *Jaw* 17	*Body* 18
u	*Lip H* 91 *Lip P* 18	*Lip H* 46 *Body* 44 *Drsm* 17	*Drsm* 10
o	*Lip H* 40 *Apex* 18 *Body* 15 *Lip P* 15	*Body* 87 *Lip H* 35	
ɔ	*Body* 23 *Lip H* 22	*Body* 73 *Jaw* 31 *Lip H* 28 *Apex* 15	

Tableau 2 : Coefficients de variation (en %), pour les gestes audibles, par voyelle, formant et paramètre.

PRÉCISIONS ET RECONNAISSANCES

L'ensemble de ce travail n'aurait pu être mené à bien sans de nombreuses collaborations et complicités scientifiques. Shinji Maeda a mis à la disposition de l'ICP son modèle à 7 paramètres ; Vincent Jacquart en a réalisé une première adaptation et implantation. Les coefficients de passage entre coupe sagittale et fonction d'aire ont été obtenus grâce au service de radiologie du CHU de Grenoble du Professeur Crouzet, et plus particulièrement avec l'aide du docteur Lebeau et de Madame Pascal-Ortiz. Le modèle du pavillon labial a été modifié avec Christian Abry. La simulation acoustique harmonique est due à Pierre Badin et Gunnar Fant, les simulations temporelles à Gang Feng (par pôles) et Éric Castelli (modèle à réflexion, SIMOND). L'ensemble de la simulation a été implanté avec HyperCard (interface HyperTalk) avec Stéphane Bernier, Patrice Vacchino et Fabien Pinet. À partir d'un dictionnaire affiné par Christophe Savariaux, les prototypes vocaliques ont été élaborés avec Nathalie Vallée. Un certain nombre de points de cette étude ont été discutés avec Christian Abry et Jean-Luc Schwartz.

RÉFÉRENCES

ABRY C. & BOË L.J. (1986), *« Laws » for Lips*, Speech Comm., 5, 97-104.

ABRY C., BOË L.J. & SCHWARTZ J.L. (1989), *Plateaus, Catastrophes and the Structuring of Vowel Systems*, J. of Phonetics, 17, 47-54.

ABRY C. & SCHWARTZ J.L. (1988/89), Présentation de *La perception visuelle de la parole : aperçu de l'état des connaissances*, M-A. Cathiard, Bulletin de l'Institut de Phonétique de Grenoble, 17-18, 109-193.

BADIN P. & FANT G. (1984), *Notes on Vocal Tract Computations*, STL QPSR, 2-3, 53-108.

BENOIT C., BOË L.J. & ABRY C. (1991), *The Effect of Context on Labiality in French*, EuroSpeech 91, 1, 153-156.

BERNIER S. & VACCHINO P. (1991), *Implantation du modèle de Maeda*, Mémoire de stage IUT2, Grenoble.

BOË L.J. & PERRIER P. (1990), *Comments on "Distinctive Regions and Modes : A New Theory of Speech Production" by M. Mrayati, R. Carré & B. Guérin* [Speech Comm. 7(3), 257-286 (1988)], Speech Comm., 9, 217-230.

BOTHOREL A., SIMON P., WIOLAND F. & ZERLING J.P. (1986), *Cinéradiographie des voyelles et des consonnes du français*, Travaux de l'Institut de Phonétique de Strasbourg.

CASTELLI E. (1988), *Caractérisation acoustique des voyelles nasales du français. Mesures, modélisation et simulation temporelle*, Thèse Docteur Ingénieur, INP Grenoble.

FENG G. (1983), *Vers une synthèse par la méthode des pôles et zéros*, 13e JEP du GALF, 155-157.

FLANAGAN J.L. (1955), *A Difference Limen for Vowel Frequency*, J. Acoust. Soc. Am., 27, 613-617.

JACQUART V. (1990), *Modélisation articulatoire du conduit vocal. Exploration et exploitation du modèle à 7 paramètres de Maeda. Étude des compensations*, DEA INP Grenoble.

LALLOUACHE M.T. (1991), *Un poste « Visage-Parole » couleur. Acquisition et traitement automatique des contours des lèvres*, Thèse de l'INP de Grenoble.

MAEDA S. (1989), *Compensatory Articulation during Speech : Evidence from the Analysis and Synthesis of Vocal-Tract Shapes using an Articulatory Model*, In *Speech Production and Modelling*, 131-149, W.J. Hardcastle & A. Marchal, (eds.), Academic Publishers, Kluwer.

MAJID R. (1986), *Modélisation articulatoire du conduit vocal. Exploration et exploitation. Fonctions de macro-sensibilité paramétriques et voyelles du français*, Thèse Docteur Ingénieur, INP, Grenoble.

MANTAKAS M. (1989), *Application du second formant effectif F'2 à l'étude de l'opposition d'arrondissement des voyelles antérieures du français*, Thèse INP Grenoble.

MERMELSTEIN P. (1978), *Difference Limens for Formant Frequencies of steady-state and Consonant-Bound vowels*, J. Acoust. Soc. Am., 63, 572-580.

MORRIS A.C. (1990), *The Use of Non-Linear Net-Input Function MLP Units for Learning The Maeda Model Function*, Internal Technical Report, ICP, Grenoble.

OHALA J. (1989), ed., *Theme Issue on the Quantal Nature of Speech*, J. of Phonetics, 17, 156 p.

PERRIER P., BOË L.J. & SOCK R., (1992), *Vocal Tract Area Function Estimation from Midsagittal Dimensions with CT Scans and a Vocal Tract Cast : Modelling the Transition with two Sets of Coefficients*, J. of Speech and Hearing Research, 35, 53-67.

RIORDAN A. (1977), *Control of Vocal Tract Length in Speech*, J. Acoust. Soc. Am., 62, 998-1002.

SLANEY M. (1990), *Interactive Signal Processing Documents*, IEEE ASSP Magazine, 7, 8-20.

STEVENS K.N. (1972), *The Quantal Nature of Speech : Evidence from the Articulatory-Acoustic Data*, In *Human Communication : A Unified View*, 51-66, E.E. David Jr. & P.B. Denes (eds.), McGraw-Hill, New-York.

STEVENS K.N. (1989), *On the Quantal Nature of Speech*, J. of Phonetics, 17, 3-45.

VALLÉE N. & BOË L.J. (1992), *Vers des prototypes articulatoires et acoustiques pour les 37 phonèmes vocaliques d'UPSID*, 19e JEP de la SFA (dans ces mêmes actes).

WOOD S. (1986), *The Acoustical Significance of Tongue, Lip, and Larynx Manœuvers in Rounded Palatal Vowels*, J. Acoust. Am., 80, 391-401.

WOOD S. (1991), *Vowel Gestures and Spectra : from Raw Data to Simulation and Applications*, XIIe Congrès International des Sciences Phonétiques, 1, 215-219.

ZERLING J.P. (1990), *Aspects articulatoires de la labialité vocalique. Contribution à la modélisation à partir de labiophotographies et films radiologiques. Étude statique, dynamique et contrastive*, Doctorat ès Lettres, Univ. des Sciences Humaines de Strasbourg.

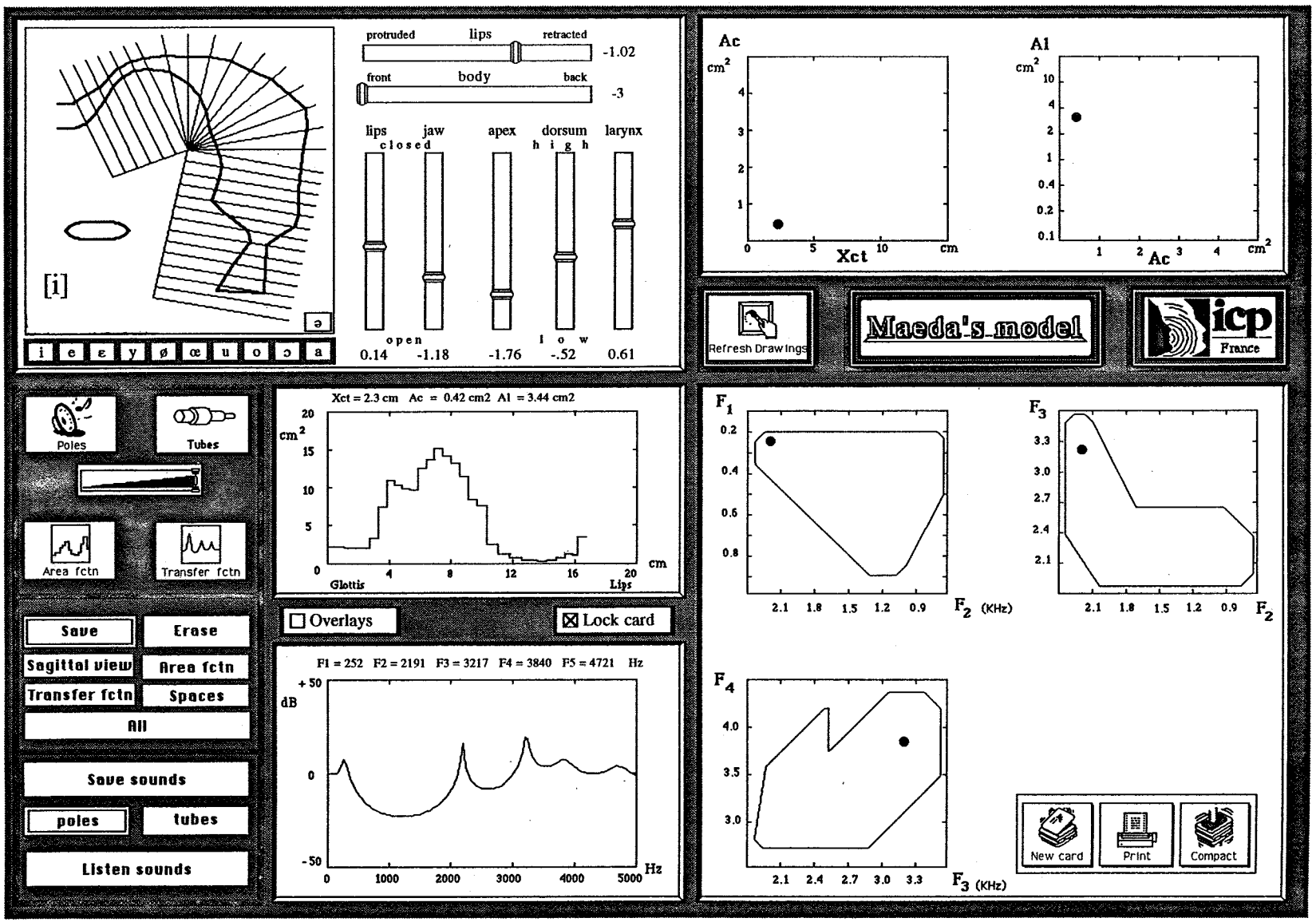


Figure 1 : Le modèle de S. MAEDA implanté sous forme de *document interactif (Interactive Modelling Document).* À partir des 7 degrés de liberté, (modifiables à l'aide de curseurs), sont calculés et affichées : la coupe sagittale, la fonction d'aire (Xct, Ac, Al correspondant au lieu de la constriction repéré par rapport aux incisives, l'aire de la constriction, l'aire aux lèvres) et la fonction de transfert (caractérisée par F1 à F4). L'item vocalique est repéré dans l'espace géométrique Xc, Ac, Al et dans l'espace maximal 4D : F1, F2, F3, F4. Il est possible de générer et d'écouter un fichier signal élaboré à partir d'une synthèse de type *pôles*, ou *tubes*.

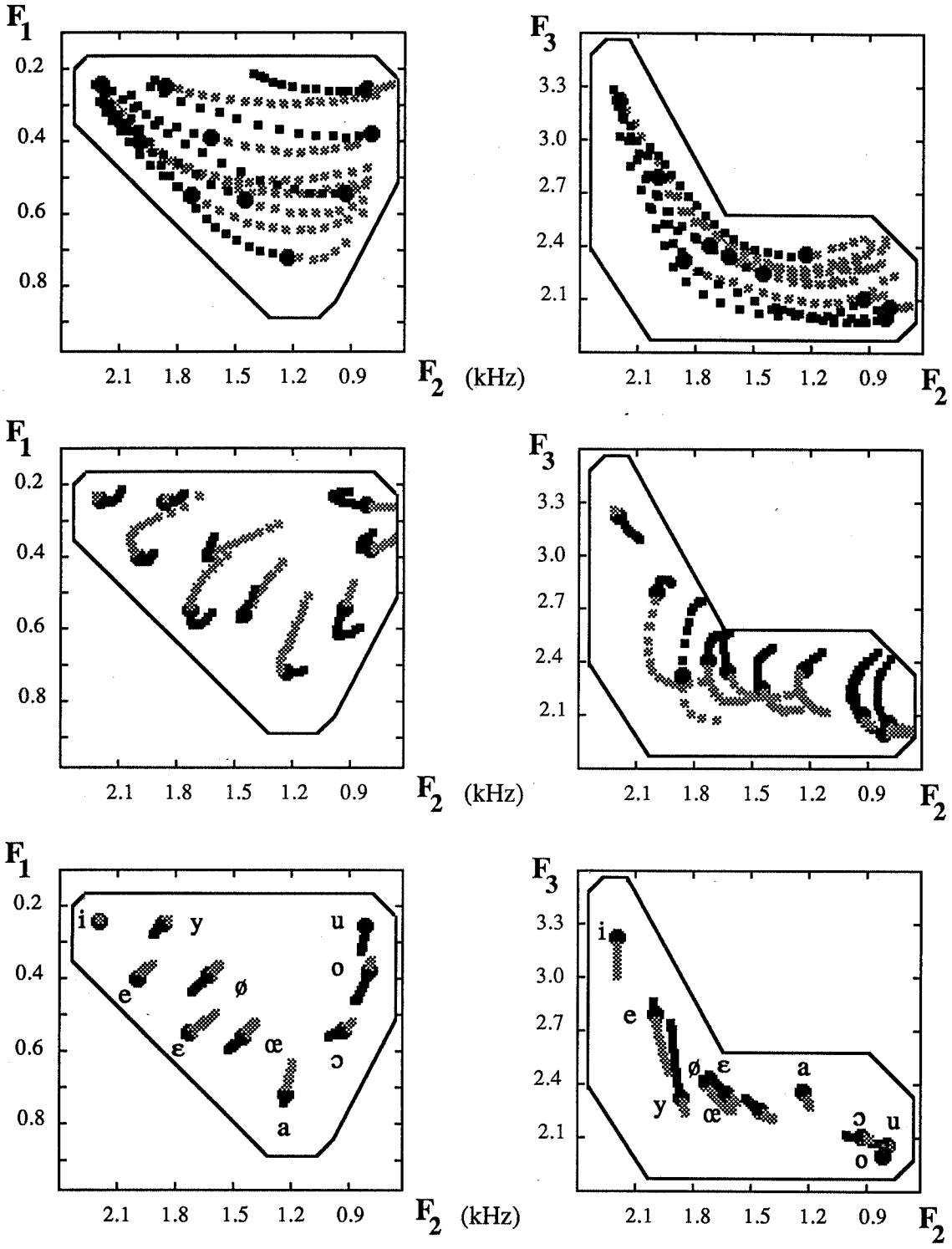


Figure 2 : Prototypes vocaliques de 10 voyelles cardinales [i, e, ɛ, a, y, ø, œ, u, o, ɔ] et les macrovariations, obtenues avec le modèle de MAEDA (1989), correspondant (de haut en bas) au Corps de la langue (*Body*), Dos de la langue (*Dorsum*) et Protrusion labiale (*Lip Protrusion*) . Exploration de ± 3 fois l'écart-type, avec un pas de 0.25 σ. Les espaces maximaux ont été tracés par exploration systématique de l'espace de commande (60 000 items vocaliques).

● Cibles vocaliques

Corps plus en avant
Dos plus abaissé
Lèvres moins protruses

Corps plus en arrière
Dos plus relevé
Lèvres plus protruses

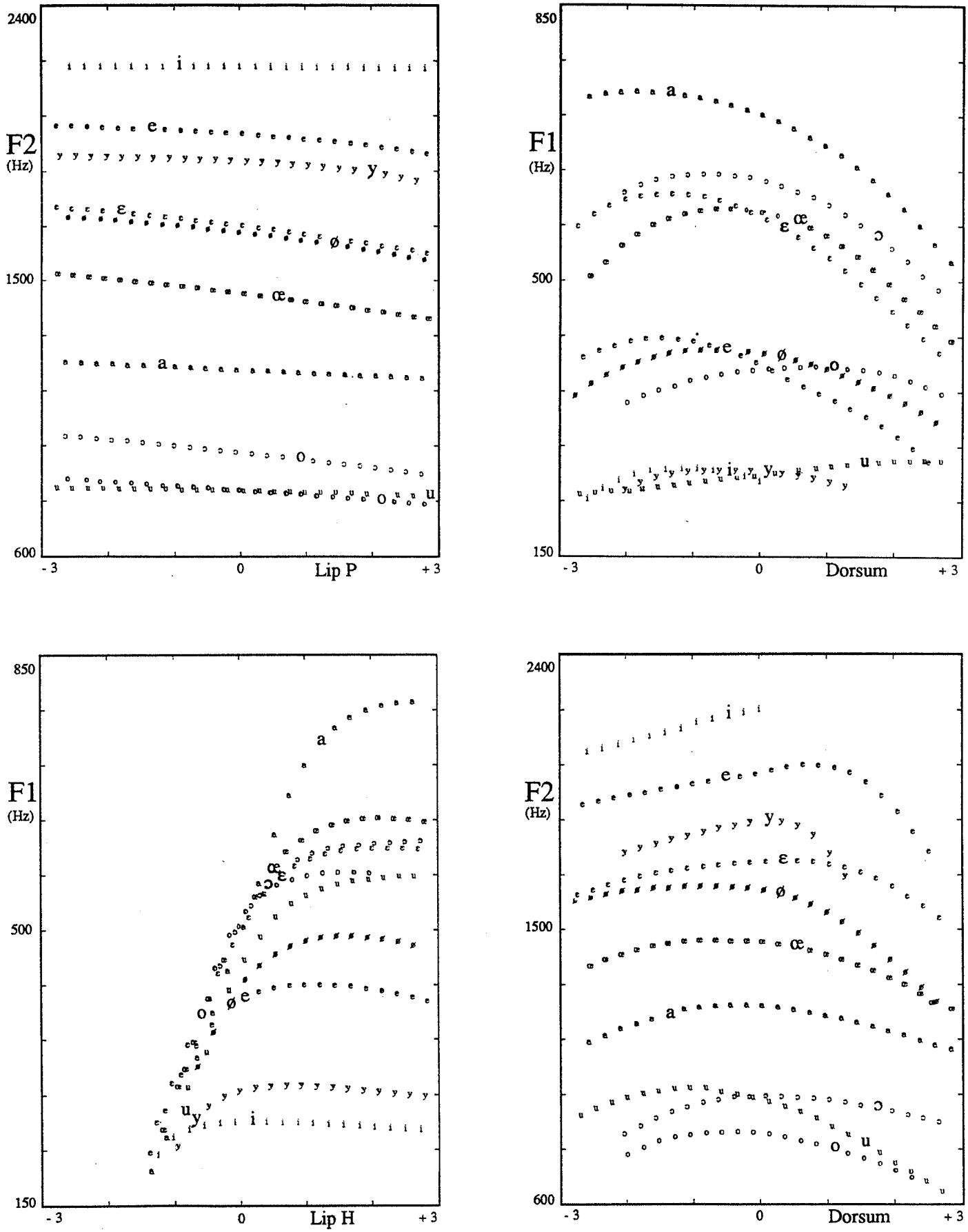


Figure 4 : Exemples de relations non-linéaires entre paramètres et formants :
- *très faible dépendance* : la protrusion labiale et F2
- *saturation* : l'écartement labial et F1
- *non-monotonie parabolique* : le dos de la langue et F1 ou F2.

DES SIGNAUX ACOUSTIQUES AUX GESTES ARTICULATOIRES : MODÉLISATION DU CONTRÔLE MOTEUR EN PRODUCTION DE LA PAROLE

Rafael LABOISSIERE, Jean-Luc SCHWARTZ et Gérard BAILLY
Institut de la Communication Parlée
CNRS URA 369 – INPG – Université Stendhal
INPG, 46 Av. Félix Viallet 38031 Grenoble Cedex
Tel : 76.57.48.49 Fax: 76.57.47.10
e-mail : rafael@icp.imag.fr

Résumé

L'objectif ultime de la phonétique est de fournir une théorie cohérente de la perception et de la production de la parole, capable d'expliquer comment des unités phonologiques discrètes sont encodées dans – et décodées de – un signal acoustique continu. Le mécanisme de transmission du code et la nature de la relation locuteur-auditeur ont fait l'objet de plusieurs théories. Cependant, ces théories ont le lourd handicap de n'avoir guère, jusqu'à présent, conduit à des modèles *quantitatifs* de la relation locuteur-auditeur. La recherche que nous poursuivons vise à combler ce manque, avec l'objectif avoué de produire des simulations quantitatives d'encodage de séquences phonologiques, des commandes motrices aux gestes articulatoires puis aux signaux acoustiques. Nous avons choisi une modélisation dynamique du contrôle moteur en production de parole s'inspirant d'un courant récent en théorie du contrôle, c'est-à-dire l'usage de réseaux neuromimétiques pour le pilotage de modèles anthropomorphiques. Une caractéristique majeure de cette approche est que les mécanismes de coarticulation y sont considérés comme le résultat d'une planification ayant pour objectif de résoudre au mieux la négociation locuteur-auditeur. Un tel cadre de modélisation permet de rendre compte de certains faits portant sur les mécanismes de coarticulation et de compensation aux perturbations. Nous montrons finalement qu' il est effectivement possible d'apprendre à *effectuer des gestes à partir des sons* et à s'engager ainsi sur la voie originale d'une *robotique de la communication parlée.*

1 INTRODUCTION

L'objectif ultime de la phonétique est de fournir une théorie cohérente de la perception et de la production de la parole, capable d'expliquer comment des unités phonologiques discrètes sont encodées dans – et décodées de – un signal acoustique continu. Ce problème alimente le débat déjà ancien sur la dialectique de l'invariance et de la variabilité (voir Perkell and Klatt, 1986). Le mécanisme de transmission du code et la nature de la relation locuteur-auditeur ont fait l'objet de plusieurs théories. Pour les tenants de la Théorie Motrice (voir Liberman et al., 1967 ; voir aussi la Théorie de la Perception Directe de Fowler, 1980), l'invariant est *articulatoire* ; cependant, sa nature exacte – configuration du conduit vocal, commande motrice à plus haut niveau – reste encore à éclaircir (voir Liberman and Mattingly, 1985). Avec la Théorie Quantique, Stevens (1972, 1989) propose au contraire une invariance *acoustique ou auditive* : les gestes articulatoires sont spécifiés par rapport à des cibles auditives, exploitant des zones de stabilité de la relation articulatori-acoutique. Lindblom (1987) quant à lui postule qu'il n'y a nulle part une réelle invariance, mais plutôt une variabilité irréductible provenant d'une négociation permanente entre locuteur et auditeur, donc d'une caractéristique essentielle de la communication parlée : sa nature *adaptative.*

Derrière cette présentation très réductrice, chacune de ces théories est alimentée par d'importants corpus de faits expérimentaux, insérée dans une argumentation forte, et traduit chez chacun de leurs auteurs un "point de vue" profond et riche d'implications sur des problèmes aussi variés que les relations phonétique-phonologie, la nature des systèmes linguistiques, le fonctionnement des systèmes perceptifs et moteurs, ... Cependant, ces théories ont le lourd handicap de n'avoir guère, jusqu'à présent, conduit à des modèles *quantitatifs* de la relation locuteur-auditeur. La recherche que nous poursuivons (Laboissière et al., 1991 ; Bailly et al., 1991) vise à combler ce manque, avec l'objectif avoué de produire des simulations quantitatives d'encodage de séquences phonologiques, des commandes motrices aux gestes articu-

latoires puis aux signaux acoustiques – et éventuellement optiques, voir Robert-Ribes et al., 1992).

Dans ce domaine, l'impressionnant travail précurseur réalisé aux Laboratoires Haskins autour de la Théorie de l'Action apparait naturellement en ligne de mire. Ces chercheurs (Kelso et al., 1986 ; Saltzman and Munhall, 1989) proposent un modèle quantitatif utilisant la notion de *structure coordinative* ou groupement fonctionnel de muscles recrutés en fonction d'une tâche donnée (Fowler, 1980). Leur modèle est caractérisé par deux points clé :

a) les cibles sont spécifiées en termes de position et degré de *constriction du conduit vocal*, et un outil mathématique sophistiqué permet de déterminer les commandes articulatoires du système de production de parole à partir des dynamiques dans l'espace des tâches ;

b) la coarticulation est considérée comme le résultat d'une *superposition spatio-temporelle linéaire* de structures coordinatives invariantes (voir Fujimura, 1991, pour des précisions sur le statut de l'hypothèse de superposition linéaire dans l'examen de la Théorie Motrice de la perception de parole).

Bien que fort élaboré, le modèle des Laboratoires Haskins, dans sa forme présente, ne peut rendre compte des mécanismes de négociation mis en avant par Lindblom, puisque précisément la cible est spécifiée par une commande invariante. D'autre part, elle ne contient pas de mécanisme de paramétrisation intégré : comment règle-t-on le timing des commandes, la "partition gestuelle" de Browman and Goldstein (1985) – voir cependant une hypothèse de travail chez Saltzman and Munhall (1989). Enfin et surtout, rien n'est dit – et peu semble pouvoir être dit – sur la manière dont les cibles spatiales (orosensorielles) pourraient être *apprises* par l'enfant dans son développement.

Afin de nous démarquer des défauts énumérés ci-dessus, nous avons choisi une modélisation dynamique du contrôle moteur en production de parole s'inspirant d'un courant récent en théorie du contrôle, c'est-à-dire l'usage de réseaux neuromimétiques pour le pilotage de modèles anthropomorphiques (Jordan, 1990 ; Jordan and Rumelhart, 1991 ; Kawato et al., 1987 ; Burnod and Dufossé, 1990). Une caractéristique majeure de cette approche est que les mécanismes de coarticulation y sont considérés comme le résultat d'une planification ayant pour objectif de résoudre au mieux la négociation locuteur-auditeur (voir Whalen, 1990 pour une expérience récente montrant l'existence de mécanismes de planification dans la production de séquences phonologiques). Nous montrerons qu'un tel cadre de modélisation permet de rendre compte de certains faits portant sur les mécanismes de coarticulation et de compensation aux perturbations. Trois problèmes importants peuvent être abordés dans ce cadre.

a) Quelle est la nature de l'espace de contrôle utilisé dans la programmation du geste ? Est-il essentiellement acoustique (notre hypothèse "minimale", comme on le verra plus loin) ou un hybride plus complexe faisant appel à des paramètres orosensoriels, voire à une action spécifique sur certains articulateurs spécialisés ?

b) Quels sont les principes adéquats de contrôle, ces principes étant définis comme le choix d'une architecture et d'un ensemble de contraintes permettant de résoudre le problème mal-posé de la spécification des gestes à partir des sons ?

c) Etant donné un modèle dynamique complet défini par les deux items précédents, est-il possible d'apprendre un modèle inverse capable de générer effectivement des commandes articulatoires acceptables à partir de la spécification des buts à atteindre (le problème de l'inversion) ?

Au-delà des questions théoriques fondamentales auxquelles nous cherchons à nous confronter – nature du contrôle, des représentations phonétiques, des relations code-signal et locuteur-auditeur – notre travail est, on le voit, un problème typique de *robotique* : étant donné un robot disposant d'un modèle des relations entre des paramètres internes configurationnels (dans notre cas, une forme du conduit vocal) et des paramètres externes positionnels (dans notre cas, des positions dans l'espace acoustique, c'est-à-dire des paramètres spectro-temporels), comment doit-on agir sur le robot afin qu'il remplisse un certain nombre de tâches dans l'espace externe ? C'est donc à la *mise en oeuvre d'un robot parlant*, capable d'apprendre à produire les *commandes adéquates* lui permettant de remplir des *objectifs acoustiques* encodant des *séquences phonologiques*, que nous nous attelons.

2 DESCRIPTION DU MODELE

Le schéma général de notre modèle est décrit sur la Fig. 1. Les lignes pleines représentent le flux des commandes et signaux durant le processus de production. Les cibles spatio-temporelles sont générées à partir des commandes phonologiques de haut niveau par le *modèle de référence*. Ces cibles spécifient la tâche à remplir. Le *modèle inverse* est en charge de la génération de patterns spatio-temporels de commande, qui fourniront finalement une sortie acoustique. Les lignes pointillées représentent le flux d'information

pendant la phase d'apprentissage. Une caractéristique essentielle de notre modèle est la présence d'un *modèle direct* supposé exister quelque part dans le système nerveux central (voir Burnod and Dufossé, 1990). Ce modèle simule en interne le passage des commandes motrices aux sorties acoustiques : il sert donc à être capable de *prévoir les conséquences de ses actions, sans passage à l'acte.* L'apprentissage de cette représentation mentale pourrait être fait au cours de la phase de babillage chez l'enfant, en exploitant le feed-back auditif.

Dans notre travail, le modèle direct consiste en une approximation analytique (par réseau neuromimétique) du modèle anthropomorphique de Maeda (1979). Ce modèle permet de réaliser le passage d'un ensemble de cinq paramètres articulatoires (mâchoire, corps de la langue, dos de la langue, pointe de la langue, fermeture/protrusion des lèvres) à la coupe sagittale du conduit vocal, puis, en appliquant la théorie acoustique, de calculer les résonances spectrales associées (soit les trois premiers formants). Nous avons ainsi préparé un ensemble de 1500 couples configuration articulatoire/configuration acoustique, couvrant le voisinage des 11 voyelles orales du français, puis nous avons approché au mieux ces données par optimisation d'un modèle neuromimétique classique, avec la technique de rétropropagation du gradient.

Une fois le modèle direct appris, il est possible d'inférer un *modèle inverse.* En effet, en partant d'une configuration articulatoire donnée et en sachant pour la sortie acoustique correspondante à quelle distance on est d'une cible donnée (spécifiée par le modèle de référence discuté précédemment), on peut estimer, dans un modèle direct *analytique*, quelles corrections apporter aux commandes articulatoires pour diminuer cette distance, donc pour réduire l'erreur. Néanmoins, le problème de l'apprentissage du modèle inverse est mal posé, étant donné l'excès de degrés de liberté de la relation commandes articulatoires/sorties acoustiques (plusieurs commandes pour le même son). Ce problème peut être régularisé par l'introduction de contraintes qui permettent la sélection d'une solution au détriment des autres. Dans notre cas, ces contraintes seront fournies par des principes d'économie articulatoire qui orientent le module de programmation (le modèle inverse) vers le choix de trajectoires aussi lisses que possibles. On voit que ce sont donc *les conditions de négociation locuteur/auditeur* qui déterminent une solution particulière pour le problème inverse.

Nous allons présenter dans la section suivante le comportement de notre modèle en relation avec deux paradigmes expérimentaux, traitant chacun de situations génératrices de variabilité (ou d'adaptabilité) pour la réalisation d'une invariance phonologique : (i) les effets de coarticulation voyelle-voyelle ou voyelle-consonne-voyelle (Öhman, 1967) et (ii) les mécanismes de compensation mis en oeuvre dans les expériences de bite-block (Lindblom et al., 1979).

3 RESULTATS

3.1 Compensation mâchoire-langue dans les expériences de bite-block

Cette expérience célèbre de Lindblom et al. (1979) a permis de montrer qu'un locuteur à qui on faisait mordre un objet de façon à le maintenir bouche ouverte était capable de compenser le blocage de sa mâchoire en position basse par un jeu accru de la langue de façon à réaliser la voyelle fermée [i]. Ces auteurs ont montré d'autre part que cette compensation était immédiate et ne passait pas par un feed-back auditif, puisque dès l'émission d'un signal, le premier formant atteignait sa valeur basse vers 250 Hz typique d'une voyelle fermée.

Notre réplication de l'expérience de Lindblom et al. s'est réalisée en deux temps. D'abord nous avons initialisé les paramètres articulatoires à une position neutre (de type [ə]) et nous les avons laissé évoluer par descente de gradient de manière à obtenir en sortie du modèle direct les valeurs formantiques typiques d'un [i] (la *situation contrôle*). Puis nous avons répété l'expérience avec la mâchoire bloquée à une position très ouverte (la *situation bite-block*). Les résultats présentés sur la Fig. 2 montrent que la cible acoustique est effectivement atteinte dans le second cas en dépit de la perturbation, grâce à une compensation mâchoire-langue. Ceci démontre la capacité de notre modèle direct à avoir appris correctement les propriétés essentielles de la transformation articulatori-acoustique, et fournit un argument fort en faveur de l'existence d'un modèle direct dans le contrôle moteur en production de parole.

3.2 Coarticulation voyelle-voyelle

Nous nous sommes intéressés dans cette expérience à l'apprentissage du modèle inverse pour la production de séquences [ua] et [ia] dans un même contexte gauche et droit de configuration neutre [ə]. Dans chaque cas nous avons spécifié la sortie acoustique par la donnée des valeurs formantiques à quatre instants correspondant respectivement aux quatre voyelles de la séquence, les trajectoires formantiques étant quant à elles laissées libres durant les transitions (voir la spécification précise Fig. 3). Les contraintes utilisées pour régulariser le problème de l'inversion étaient fournies par l'introduction d'un coût articulatoire à minimiser (voir Laboissière et al., 1991 pour une description précise de l'implémentation). Nos résultats (comparer les configurations du conduit vocal pour

[u] en contexte droit [a], Fig. 4a, vs [i], Fig. 4b) montrent que nous avons effectivement obtenu de la coarticulation : durant la production de [u], la langue anticipe – autant que l'on peut se le permettre pour réaliser [u] – la voyelle suivante, avec donc une position plus avancée en contexte [i] qu'en contexte [a], alors que la cible acoustique de [u] est correctement atteinte dans les deux cas.

3.3 Apprentissage de gestes consonantiques

Un test critique de notre modèle est la possibilité dans un tel cadre de simulation d'inférer à partir de signaux acoustiques les gestes articulatoires corrects pour la production des consonnes. Nous appelons gestes articulatoires corrects des gestes qui (a) recrutent les bons articulateurs (c'est-à-dire la pointe de la langue pour les dentales, le dos de la langue pour les vélaires, les lèvres pour les labiales), et (b) présentent les bons patrons de coarticulation, c'est-à-dire respectent ce que l'on sait des lieux de constriction pour les plosives, par exemple.

Ce problème est bien sûr considérable, et presque entièrement à traiter. Nous voulons présenter ici quelques données préliminaires sur l'apprentissage de gestes de la langue pour la production de plosives, c'est-à-dire de gestes [d] et [g] en contexte [a#a]. La voyelle était spécifiée dans cette expérience par des cibles formantiques comme dans le cas précédent, tandis que la consonne était spécifiée par son "locus" dans le domaine spectral. De façon à obtenir une réelle fermeture, nous avons ajouté à notre modèle direct un module d'estimation du degré de constriction du conduit vocal à partir des paramètres de commande, et nous avons imposé une fermeture totale pour la consonne. La spécification de la tâche, ainsi que les résultats obtenus, sont présentés dans la Fig. 5.

Ces résultats sont plutôt satisfaisants. En effet, on constate que le modèle inverse a pu inférer de la demande acoustique une stratégie de commande tout à fait acceptable, avec une avancée du corps de la lague et une montée de la pointe de la langue pour la production de la dentale [d], et une montée du dos de la langue pour la production de la vélaire [g], qui présente un lieu de constriction relativement arrière, ce qui est normal en contexte [a], mais pas pharyngal comme l'obtiennent Saltzman and Munhall (1989) avec une spécification explicite dans un espace orosensoriel.

4 Conclusion

Les résultats présentés ici nous semblent tout à fait prometteurs. Ils montrent que l'on peut effectivement apprendre à piloter un modèle anthropomorphique de production de la parole pour encoder des séquences phonologiques et produire de signaux acoustiques, et à reproduire jusqu'à un certain point les mécanismes cruciaux que sont coarticulation et compensation. Ce type de modélisation inspirée de concepts classiques de robotique et d'outils connexionnistes puissants nous permet d'avancer dans le domaine de la synthèse articulatoire vers la réalisation d'un véritable robot parlant.

Il reste que notre travail doit maintenant se confronter à deux séries de questions cruciales pour les travaux sur la commnication parlée.

a) Quels sont les principes de contrôle, les variables contrôlées, la nature de la négociation locuteur-auditeur ? Notre cadre de modélisation doit fournir un outil quantitatif permettant d'avancer sur ces problèmes.

b) Notre travail est organisé autour d'une hypothèse centrale, hypothèse "minimale", selon laquelle il est possible d'apprendre à effectuer des gestes à partir des sons (et éventuellement des images). nCeci pose la question de la capacité du système perceptif à estimer les paramètres acoustiques nécessaires (*problème de l'estimation perceptive*) et à les mettre en relation avec l'espace acoustique propre du modèle direct (*problème de la normalisation*). Ces questions sont parfois particulièrement difficiles, mais elles ne sont pas sans solution (voir Schwartz et al., 1991), et elles montrent en tout état de cause un intérêt de notre démarche, qui est de remettre au coeur des recherches en communication parlée le problème, peu considéré en général, des interactions perception-action, dont l'importance est plus largement reconnue dans d'autres domaines des Sciences Cognitives.

Remerciements

Le premier auteur est titulaire d'une bourse du Conseil National de Recherche et Développement (CNPq), Brésil, dans le cadre de l'accord FIAS-CNPq (numéro de dossier 92.0208/88.6). Il est aussi à l'Institut Technologique d'Aéronautique (ITA–CTA), São José dos Campos, Brésil.

Références

Bailly, G., Laboissière, R. et Schwartz, J.L. (1991). Formant trajectories as audible gestures: an alternative for speech synthesis. *Journal of Phonetics*, **19**(1) : 9–23.

Browman, C.P. et Goldstein, L.M. (1985). Dynamic modeling of phonetic structure. In Fromkin, V.A. (Ed.), *Phonetic Linguistics*, pp. 35–53. Academic Press, Inc.

Burnod, Y. et Dufossé, M. (1990). A model for the cooperation between cerebral cortex and cerebellar cortex in movement learning. In Paillard, J. (Ed.), *Brain and Space*. Oxford University Press.

Fowler, C.A. (1980). Coarticulation and theories of extrinsic timing control. *Journal of Phonetics*, **8** : 113–133.

Fujimura, O. (1991). Beyond the segment. In Mattingly, I.G. et Studdert-Kennedy, M. (Eds.), *Modularity and the Motor Theory of Speech Perception*, pp. 25–31. Erlbaum.

Jordan, M.I. (1990). Motor learning and the degrees of freedom problem. In Jeannerod, M. (Ed.), *Attention and Performance*, ch. XIII. Hillsdale, NJ: Erlbaum.

Jordan, M.I. et Rumelhart, D.E. (1991). Forward models: Supervised learning with a distal teacher. *(Submitted to Cognitive Science)*.

Kawato, M., Furukawa, K. et Suzuki, R. (1987). A hierarchical neural-network model for control and learning of voluntary movement. *Biological Cybernetics*, **57** : 169–185.

Kelso, J.A.S., Saltzman, E.L. et Tuller, B. (1986). The dynamical theory on speech production: Data and theory. *Journal of Phonetics*, **14** : 29–60.

Laboissière, R., Schwartz, J.L. et Bailly, G. (1991). Motor control for speech skills: A connectionist approach. In Touretzky, D.S., Elman, J.L., Sejnowski, T.J. et Hinton, G.E. (Eds.), *Proceedings of the 1990 Connectionist Models Summer School*, San Mateo, CA, pp. 319–327.

Liberman, A.M., Cooper, F.S., Schankweiler, D.P. et Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review*, **74** : 431–461.

Liberman, A.M. et Mattingly, I.G. (1985). The motor theory of speech perception revisited. *Cognition*, **21** : 1–36.

Lindblom, B. (1987). Adaptive variability and absolute constancy in speech signals: two themes in the quest for phonetic invariance. In *Proceedings of the XIth International Congress of Phonetic Sciences*, Tallin, Estonia, volume 3, pp. 9–18.

Lindblom, B., Lubker, J. et Gay, T. (1979). Formant frequencies of some fixed-mandible vowels and a model of speech-motor programming by predictive simulation. *Journal of Phonetics*, **7** : 141–161.

Maeda, S. (1979). An articulatory model of the tongue based on a statistical analysis. *J. Acoust. Soc. Am.*, **65** : S22.

Öhman, S. (1967). Numerical model of coarticulation. *J. Acoust. Soc. Am.*, **41** : 310–320.

Perkell, J.S. et Klatt, D.H. (Eds.) (1986). *Invariance and Variability in Speech Processes*. Hillsdale, NJ: Erlbaum.

Robert-Ribes, J., Escudier, P. et Schwartz, J.L. (1992). Modèles d'intégration audition-vision dans la perception des voyelles: une étude neuromimétique. In *Actes du Cinquième Colloque de l'ARC*. Nancy, France.

Saltzman, E.L. et Munhall, K.G. (1989). A dynamical approach to gestural patterning in speech production. *Ecological Psychology*, **1**(4) : 1615–1623.

Schwartz, J.L., Berthomier, F. et Escudier, P. (1991). Architectures auditives pour le décodage acoustico-phonétique. In Caelen, J., Abry, C. et Schwartz, J.L. (Eds.), *Cognition, Perception et Action dans la Communication Parlée*. (à paraître).

Stevens, K.N. (1972). The quantal nature of speech: Evidence from articulatory-acoustic data. In David Jr., E.E. et Denes, P.B. (Eds.), *Human Communication: A unified view*, pp. 51–66. New York: McGraw-Hill.

Stevens, K.N. (1989). On the quantal nature of speech. *Journal of Phonetics*, **17** : 3–45.

Whalen, D.H. (1990). Coarticulation is largely planned. *Journal of Phonetics*, **18**(1) : 3–35.

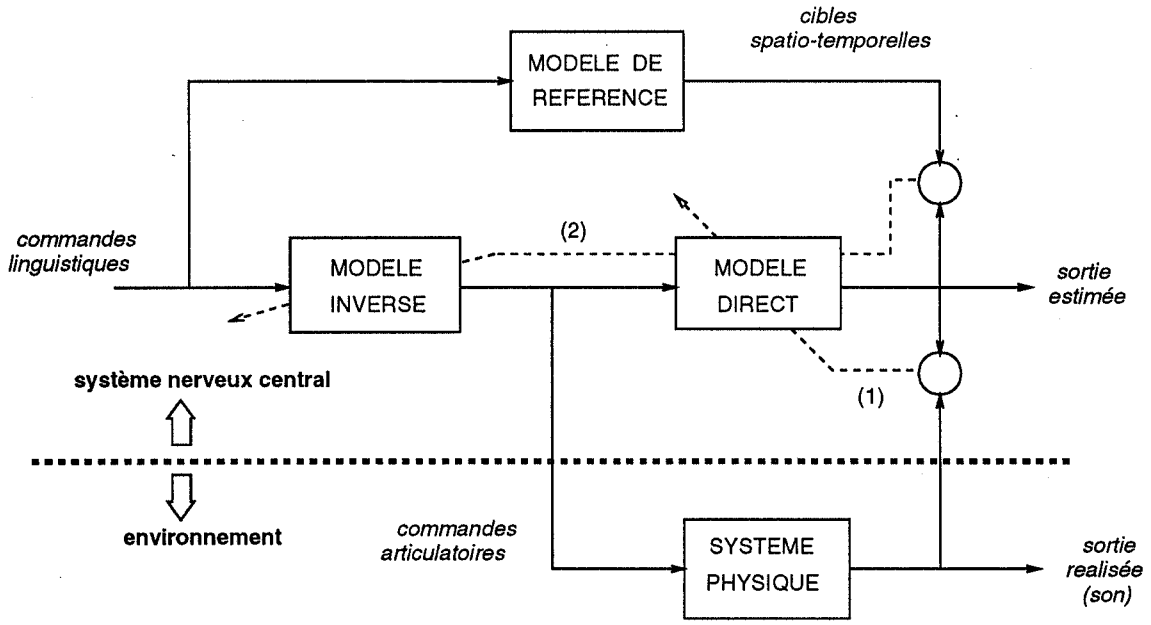


Figure 1: Cadre général de modélisation. Les traits pleins représentent le flux d'information dans le système. Les traits pointillés représentent les signaux d'erreur utilisés lors de l'apprentissage du modèle direct à travers le feed-back auditif (1) et du modèle inverse à travers le feed-back "interne" (2).

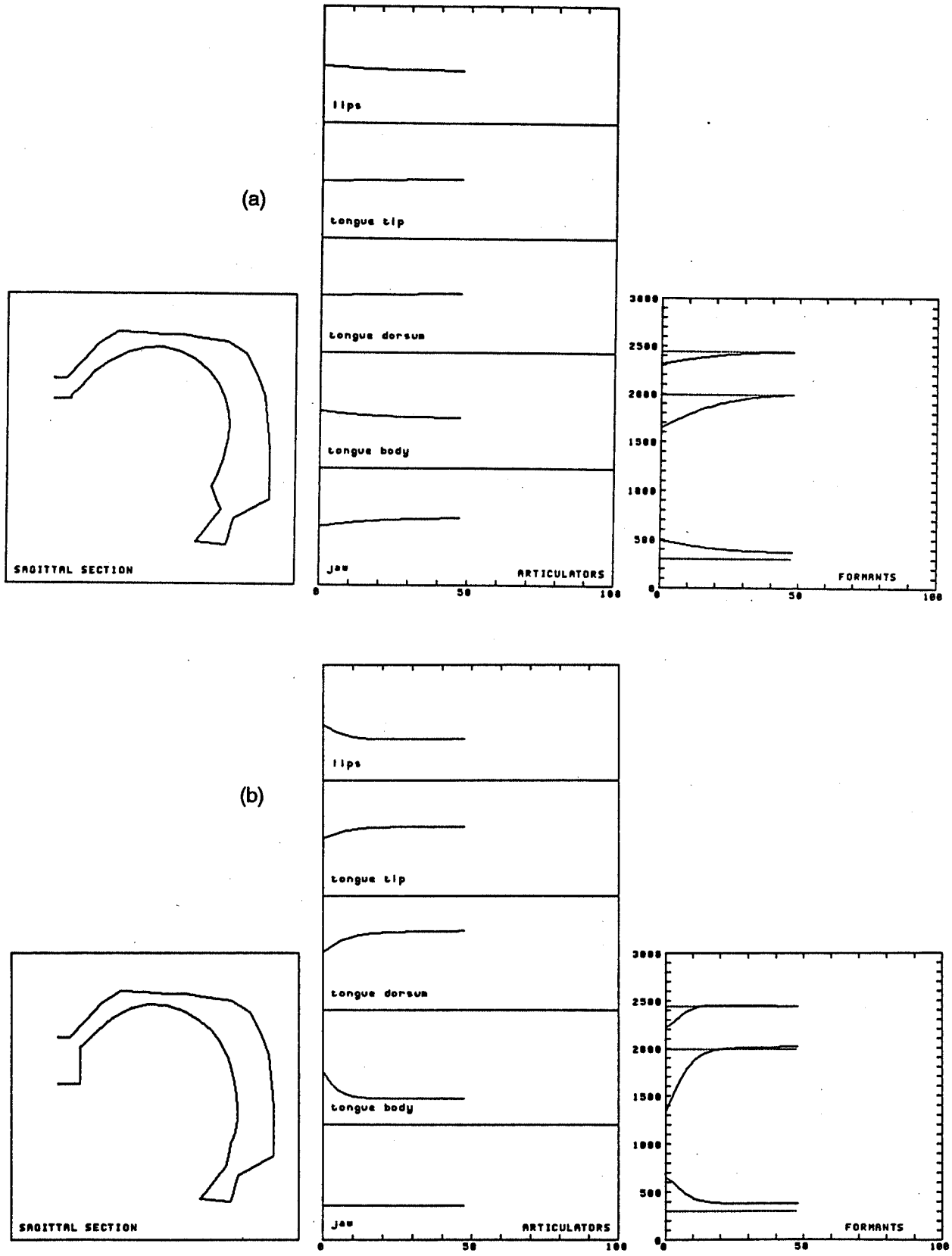


Figure 2: Expérience de "bite-block" pour la voyelle /i/ : condition normale (a) et condition bite-block (b), c.-à-d. le paramètre mâchoire est bloqué à une valeur correspondante à une configuration ouverte. A gauche : coupe-sagittale à la fin du processus d'apprentissage. A droite : évolution des trois premiers formants lors des itérations de descente du gradient ; les formants cibles sont montrés en pointillé. Au centre : évolution des paramètres articulatoires.

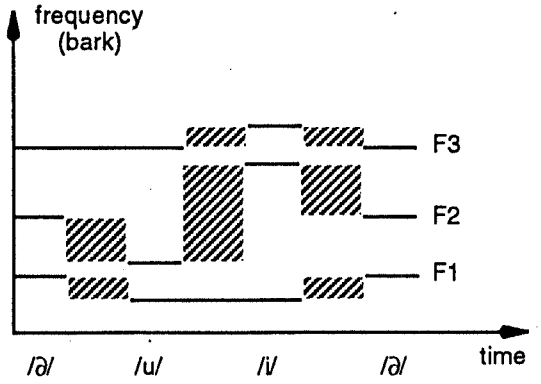
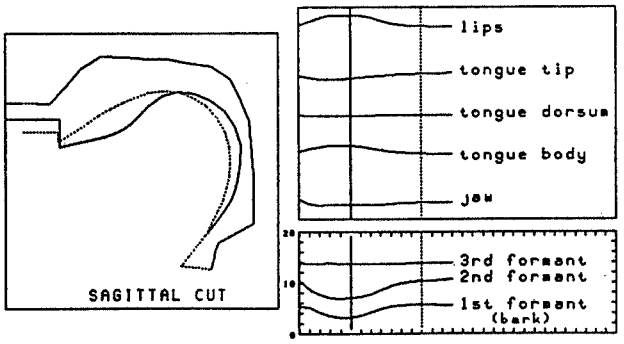
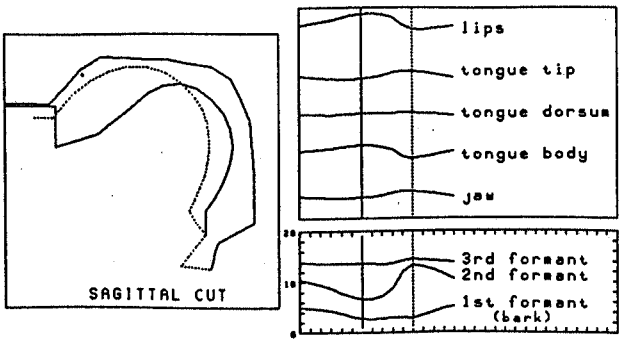


Figure 3: Spécification des cibles acoustiques (trois premiers formants) en sortie du modèle direct au cours du temps. La séquence montrée correspond à une suite phonologique /əuiə/. Les zones hachurées correspondent aux transitions formantiques entre les noyaux stables des voyelles (traits pleins).



(a)



(b)

Figure 4: Coarticulation de la voyelle /u/ dans deux contextes vocaliques différents : voyelle /u/ (traits pleins) et soit /a/ (a) ou /i/ (b) (traits pointillés). En haut à droite : trajectoires articulatoires ; en bas à droite : trajectoires formantiques ; les lignes verticales correspondent aux instants d'échantillonnage de la coupe sagittale.

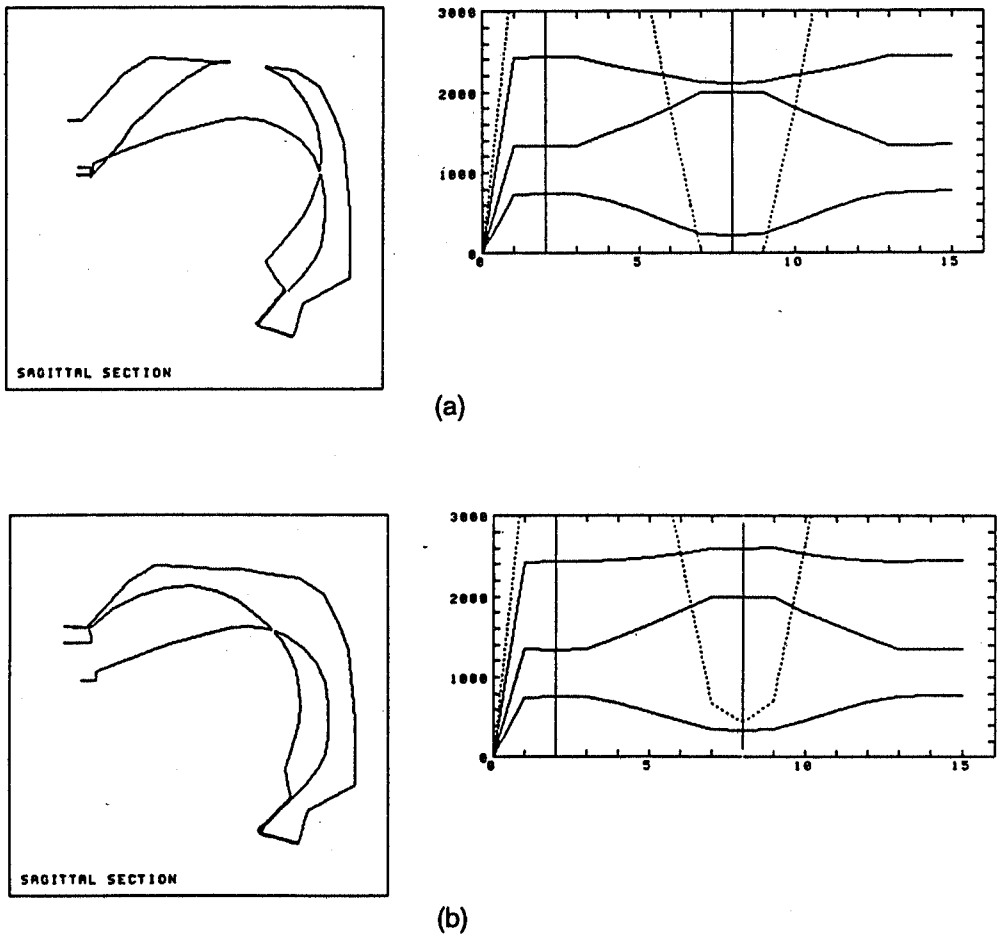


Figure 5: Résultats de la simulation de deux gestes consonnantiques dans un contexte /a#a/ : /aga/ (a) ou /ada/ (b). A gauche : coupe sagittale à deux instants différents, dans la partie stable de la voyelle initiale /a/ et à l'occlusion pour /g/ (a) et /d/ (b). A droite : trajectoires des trois premiers formants en échelle Hertz ; les deux lignes pleines verticales correspondent aux instants déchantillonnage du côté gauche. La sortie du modèle direct qui calcule de degré de constriction dans le conduit vocal est montrée sur le côté droit (traits pointillés).

19es J.E.P. - Bruxelles, 19 au 22 mai 1992.

TRANSITIONS FORMANTIQUES CORRESPONDANT A DES CONSTRICTIONS REALISEES DANS LA PARTIE ARRIERE DU CONDUIT VOCAL.

O. AL DAKKAK, M. MRAYATI,
INSTITUT SUPERIEUR DES SCIENCES APPLIQUEES ET DE TECHNOLOGIE,
BP 7028, DAMAS, SYRIE.

R. CARRE,
DEPARTEMENT SIGNAL, ENST, UNITE ASSOCIEE AU CNRS,
46 RUE BARRAULT, PARIS.

Résumé

De nombreuses langues exploitent des phonèmes réalisés par des constrictions placées dans la partie arrière du conduit vocal. On se propose d'expliquer les caractéristiques acoustiques de ces réalisations, et tout particulièrement les transitions formantiques, dans le cas des consonnes oropharyngales, pharyngales et glottales arabes, au moyen du concept de régions distinctives (4).

Nous présentons des données sur la parole naturelle, analysées par sonagraphe. Ceci nous permet de compléter le tableau classique des transitions formantiques, pour des constrictions réalisées dans la partie avant du conduit vocal [6]. Les logatomes prononcés sont de la forme /VCV/ où V=/ɛ/ et C balaie les consonnes oropharyngales, pharyngales et glottales de l'arabe.

I. INTRODUCTION

La production de la parole est un processus bien complexe. Le signal acoustique qui part du locuteur vers l'auditeur, est le résultat de plusieurs événements se passant au même temps:

- un changement dynamique dans la configuration du conduit vocal du locuteur réalisé par ses articulateurs,
- un changement résultant dans le signal acoustique, de ses paramètres spectraux et des trajectoires formantiques associées
- une perception par l'auditeur des différents sons produits...etc.

La théorie des régions distinctives présente une vue globale unifiée de ces événements [1,4,5]. Partant d'une configuration quelconque du conduit vocal, on peut passer à toute autre configuration par une superposition non-linéaire de configurations primitives. Les configurations primitives correspondent à un tube acoustique uniforme, ayant une seule constriction à une région. En dynamique, cette constriction donne lieu à des transitions formantiques bien définies par cette théorie.

Des données sur les transitions formantiques sont bien décrites, dans la littérature, pour les sons qui se produisent par constrictions dans la partie avant du conduit vocal [3,6]. Quant aux constrictions produites à la partie arrière du conduit vocal, elles sont très peu montrées. Récemment, Carré et Mrayati [1], ont étudié les trajectoires formantiques dans l'optique de la théorie des régions distinctives.

Nous présentons ici des résultats sur la parole naturelle, concernant la partie arrière du conduit vocal. Nous montrons les sonagrammes des consonnes oropharyngales, pharyngales et glottales arabes. Ces phonèmes particuliers de l'arabe ont été choisis, parce qu'ils représentent des cas de transitions formantiques lors de la production des consonnes dans la partie arrière du conduit.

En effectuant une constriction consonnantique à partir d'une configuration quasi neutre du conduit vocal, on obtient des logatomes de la forme VCV où V est la voyelle neutre /ə/ et C balaie les consonnes oropharyngales, pharyngeales et glottales de l'arabe. Dans le cas des consonnes glottales, aucune constriction n'est observée dans le conduit.

II. LES TRANSITIONS FORMANTIQUES DANS LA PARTIE ARRIERE DU CONDUIT VOCAL:

Dans la théorie des régions distinctives, (D.R.M), on divise le conduit vocal en 8 régions - si on se contente des trois premiers formants- . Les longueurs de ces régions sont : L/10, L/15, 2L/15, L/5, L/5, 2L/15, L/15 et L/10. ou L est la longueur effective du conduit vocal . Ces régions sont appelées A, B, C, D, D*, C*, B* et A*. en partant des lèvres.

Si on effectue une constriction non complète à l'une des régions, tout en restant dans le mode d'un seul tract (one tract mode OTM), et ceci à partir d'un conduit vocal quasi-cylindrique ; les transitions formantiques des trois premiers formants sont bien décrites [4] (fig. 1).

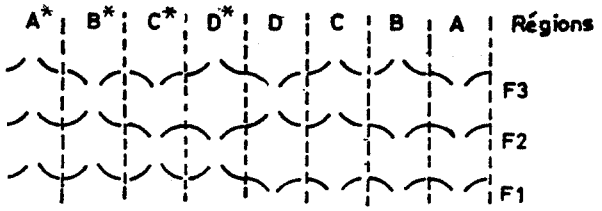
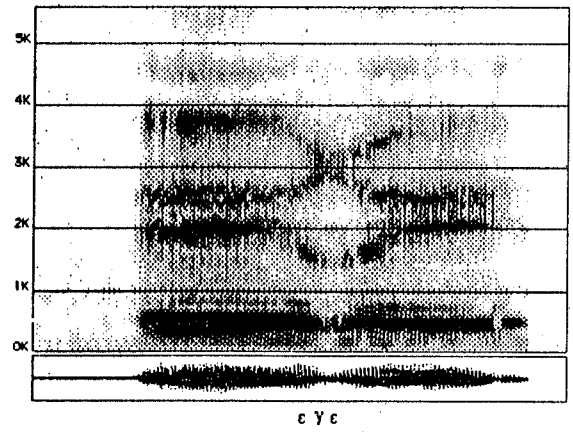
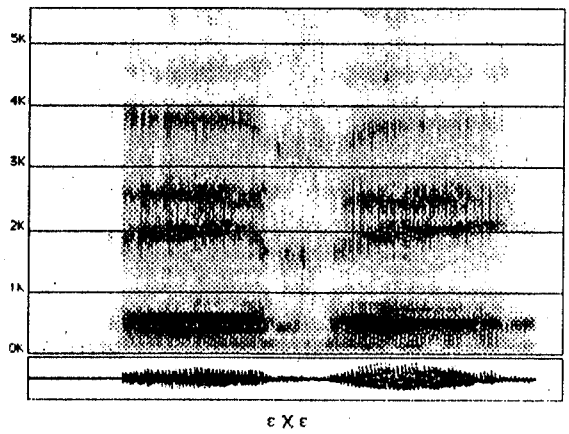


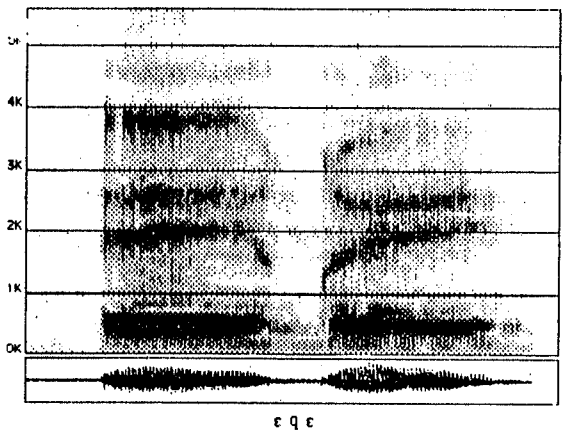
Fig .1. Les 8 transitions formantiques primitives correspondant aux 8 régions du conduit vocal

En effet, pour approcher un conduit vocal quasi-cylindrique, il faut prononcer la voyelle neutre /ə/ du shwa. Or, cette voyelle est, en général, mal prononcée par les locuteurs arabes. A la place, on les a fait prononcer le /ɛ/ courant dans la dialecte de Damas. Cette voyelle donne une configuration proche de celle donnée par le shwa. Les consonnes utilisées sont les consonnes oropharyngales et pharyngales, qui donnent des constrictions dans la partie arrière du conduit vocal. Ces consonnes sont produites dans les régions D* et C* du modèle DRM, les transitions formantiques sont anti-symétriques à celles produites par des consonnes des régions D et C respectivement. Notamment, la transition montante de F1 est bien distinctive de cette partie (fig. 2) elle est remarquable, particulièrement sur les sonagrammes de /ʕ/ et /ʁ/.

De plus, on a fait prononcer les consonnes glottales /h/, /ʔ/. Le conduit vocal ne subit aucune constriction dans les régions. On n'a aucune transition formantique.







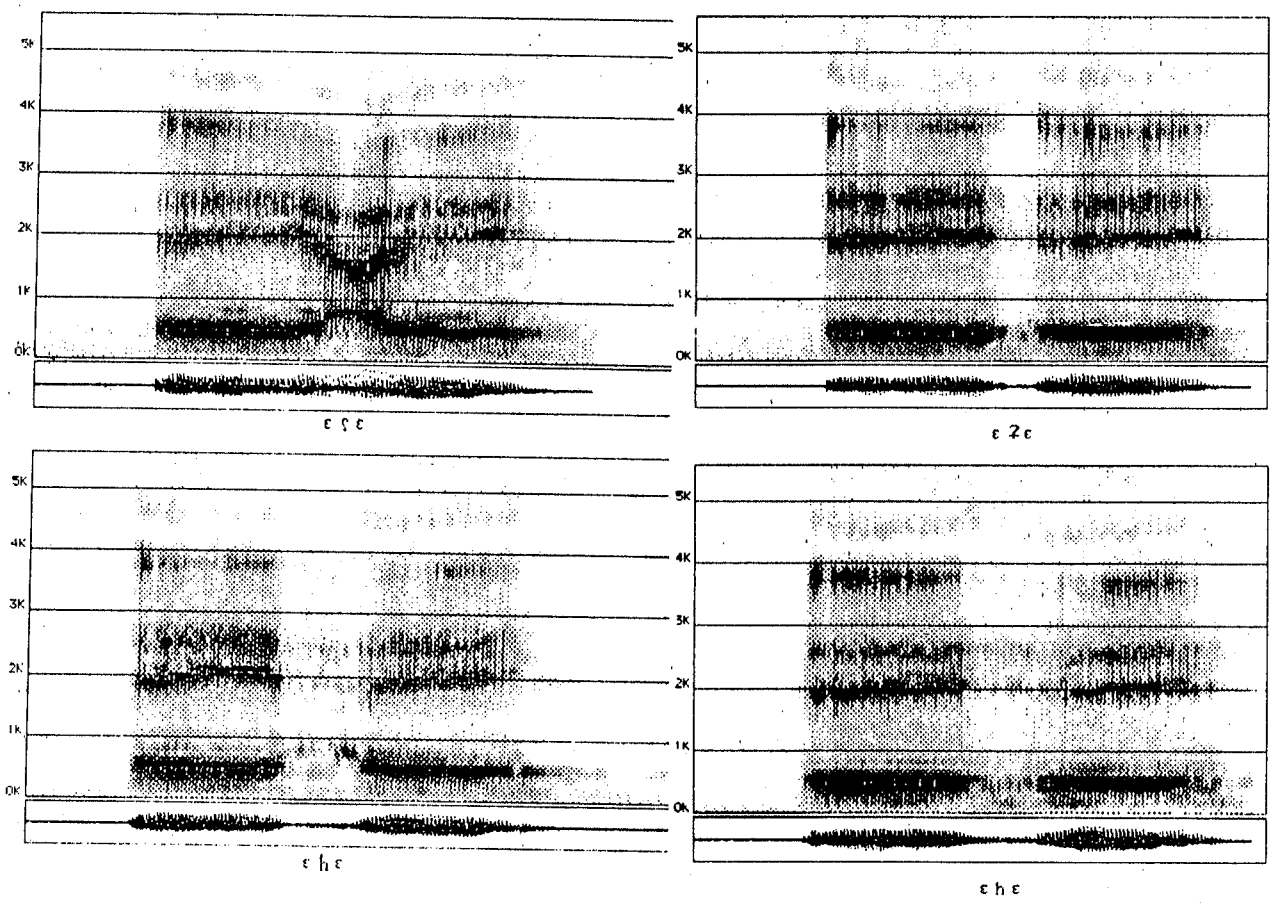


Fig. 2. Sonagrammes des logatomes/ɛCɛ/. C balaie les 7 consonnes arabes produites dans la partie arrière du CV

III. DESCRIPTION DES CONSONNES OROPHARYNGALES, PHARYNGALES ET GLOTTALES DE L'ARABE:

Dans la langue arabe, il existe sept consonnes produites dans la partie arrière du conduit vocal, qui sont les suivantes:

- /ʕ/: une consonne voisée, pharyngale, réalisée par une constriction dans la région C*

/ħ/: une fricative non voisée, pharyngale, réalisée par une constriction dans la région C*

-/ɣ/:une fricative voisée, oropharyngale, réalisée par une constriction dans la région D*. Il ressemble au /r/ voisé dans le mot francais (paris)

-/X/:une fricative non voisée, oropharyngale, réalisée par une constriction dans la région D*. Il ressemble au /r/ non voisée dans le mot francais (crier)

-/q/:une occlusive voisée, oropharyngale (uvélaire) réalisée dans la région D*.

On remarque l'absence des consonnes réalisables par des constrictions aux régions A*, B*. En effet, on a peu ou pas de contrôle sur ces régions. De plus, le volume de la cavité du pharynx est insuffisant pour permettre le murmure et la production des plosives pharyngales dans A* ou B*.

Les consonnes glottales arabes sont au nombre de deux :

-/h/: la consonne aspirée du /h/ dans le mot anglais (hat).

-/ʔ/:la consonne qui se produit quand on prononce un /a/ au début de parole, exemple dans le mot français (avant).

IV EXPERIMENTATION

Nous avons fait prononcer le corpus VCV décrit ci-dessus (V= /ɛ/; C une consonne arabe oropharyngale, pharyngale ou glottale) par 7 locuteurs de Damas. Nous avons enregistré les données sur Macintoch, en utilisant l'outil 'Sound Edit', qui permet de digitaliser la parole enregistrée sur 8 bits, et d'analyser les signaux par sonagraphe digital et par bien d'autres méthodes d'analyse spectrale. Les signaux ont été enregistrés avec une fréquence d'échantillonnage de 11 KHz (le plus proche de 10 KHz permis par le logiciel.).

Ensuite une analyse par sonagraphe large-bande avec des filtres de 128 points a été effectuée. Des exemples de sonagrammes sont présentés ci-après.

Sur ces sonagrammes, nous avons remarqué la consistance des résultats avec la prévision de la théorie de DRM pour les consonnes oropharyngales et pharyngales produites dans les régions C* et D*. Un seul locuteur avait tendance à fermer ses lèvres lors de l'enregistrement; son troisième formant est très souvent descendant.

Nous présentons aussi un exemple de sonagrammes des deux consonnes glottales de l'arabe du même locuteur; aucune transition n'est remarquée. En effet, pour ces consonnes, aucune région du conduit vocal ne subit de changement d'aire.

Il est intéressant de remarquer que ces résultats, combinés avec d'autres connus dans la littérature [7, 8, 9], mettent en évidence les liens entre les lieux d'articulation et les transitions formantiques. Si l'on considère la production de /əCə/, avec une seule constriction à la fois, le lieu d'articulation de C est clairement correlé avec les transitions formantiques. La théorie de régions distinctives (DRM), formalise cette corrélation (Fig. 3). Elle explique aussi les transitions formantiques produites, lorsque une autre voyelle que le /ə/ est prononcée avec le phénomène de coarticulation [1].

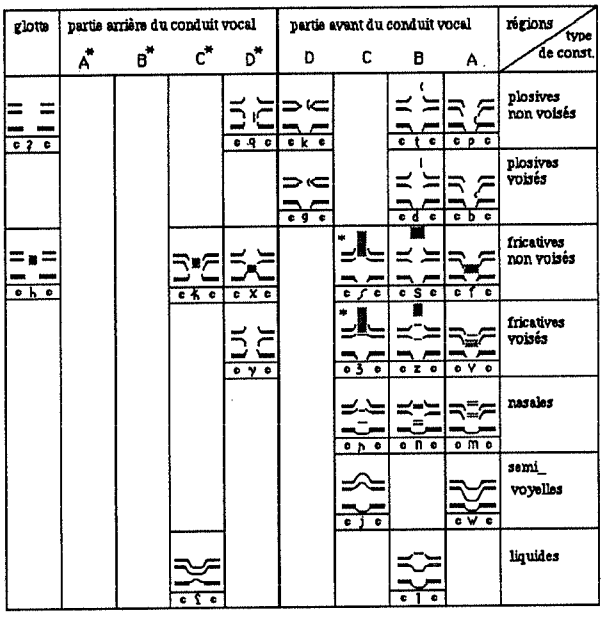


Fig. 3 Le tableau complet des transitions formantiques pour les constrictions aux différentes régions du conduit vocal

* sur la parole naturelle, tendance à fermer les lèvres, F3 descend

V. CONCLUSION

La langue arabe est l'une des langues qui contient des consonnes produites dans la partie arrière du conduit vocal. Les transitions formantiques de ces consonnes sont bien expliquables par la théorie des régions distinctives. En ajoutant nos résultats à ceux connus dans la littérature sur les transitions formantiques dans la partie avant du conduit vocal, on aurait fait un pas pour compléter l'image sur les transitions formantiques (fig. 3). Des données sur la parole réelle de ces consonnes sont présentées pour la première fois. Une étude sur les autres consonnes arabes va être effectuée ultérieurement.

BIBLIOGRAPHIE

[1] Carré, R., Mrayati, M. (1990), "Articulatory-acoustic-phonetic relations and modeling", in Speech production and speech modeling, (W.J. Hardcastle and A. Marchal, eds.), Kluwer Academic Publishers.

[2] Carré, R., Mrayati, M. (1991), "Vowel-vowel trajectories and région modeling", Journal of Phonetics, 19, 433-443.

[3] Delattre, P. (1967), "Des indices acoustiques aux traits pértinents", Proc. of the 6th ICPhS, 35-46.

[4] Mrayati, R., Carré, R., and Guérin, B. (1988), "Distinctive régions and mode: a new theory of the speech production", Speech Communication, 7, 257-286.

[5] Mrayati, M., and Carré, R. (1991), "Static and dynamic relations between vocal tract configurations and acoustics", Proc. of the 12th ICPhS.

[6] Ohman, S.E.G. (1966), "Coarticulation in VCV utterances: spectrographic measurements", J. of Acoust. Soc. of Am., 27, 484-493.

[7] Stevens K. N. (1980), "Acoustic Correlates of some Phonetic Categories", J. Acoust. Soc. Am. 68(3), 836-842.

[8] Delattre P. C., Liberman A. M., Cooper F. S.(1962) "Formant Transitions and Loci as Acoustic Correlates of Place of Articulation in American Fricatives" Haskins lab. 104-121.

[9] Sharf D. J., Hemeyer T. (1972) "Identification of Place of Consonant Articulation from Vowel Formant Transitions", J. Acoust. Soc. Am. 51, 652-658.

19es J.E.P. - Bruxelles, 19 au 22 mai 1992.

ANALYSE ACOUSTIQUE, PERCEPTIVE ET FONCTIONNELLE DES HESITATIONS VOCALES EN PAROLE SPONTANEE

ISABELLE GUAïTELLA

Laboratoire Parole et Langage URA CNRS 261
29 av. R. Schuman 13621 Aix-en-Provence

Résumé

Nous proposons une analyse des hésitations vocales à travers les paramètres de durée segmentale et de fréquence fondamentale. Cette analyse nous permet de justifier leur rôle au sein du système de fonctionnement de la prosodie en parole spontanée, système qui prend en compte les capacités de production et de perception du locuteur.

1- INTRODUCTION: CADRE THEORIQUE, HYPOTHESES ET OBJECTIFS

"alors tu vois j'ai choisi un extrait de parole spontanée où il n'y avait pas trop d'hésitations..."

Les hésitations sonores en parole spontanée constituent le plus souvent la cible privilégiée des jugements normatifs portés à l'encontre de ce type de parole, y compris par ceux qui prétendent l'étudier. A l'impression générale selon laquelle ce sont *les autres* qui hésitent, se surajoute l'idée que l'hésitation est inutile et nuisible à la communication, et qu'il est donc nécessaire de l'éliminer. On ne saurait en aucun cas l'étudier...

Cependant l'objet de cette étude est de montrer:

- que l'hésitation vocale mérite d'être étudiée, notamment sur le plan acoustico-phonétique, au même titre que tout phénomène sonore produit par le locuteur,

- que rien n'autorise à "couper" le signal de parole afin de supprimer les éléments considérés, à-priori, comme gênants et dont fait partie l'hésitation vocale.

-et que, sa présence se manifestant dans toute occurrence de parole spontanée, elle est nécessaire à la production mais aussi à la perception de la parole.

Si les pauses silencieuses ont été étudiées comme le moment privilégié de planification du message (Goldman-Eisler 1958, 1968, 1972; Butterworth, Goldman-Eisler 1979), il ne s'agit, d'après Butterworth (1980), que d'un postulat méthodologique. En effet, rien ne permet d'affirmer qu'en parole spontanée les pauses silencieuses sont uniquement là pour permettre au locuteur de préparer son message. En d'autres termes, les pauses silencieuses peuvent avoir d'autres fonctions y compris en parole spontanée (Autesserre, Nishinuma, Guaïtella 1989; Davis, Léon 1989; Guaïtella, Autesserre, Nishinuma 1990; Guaïtella, Santi 1990 + à paraître). Il est alors difficile de déterminer la ou les fonction(s) d'une pause. Parmi les différents procédés d'expression qui peuvent manifester la planification du propos par le locuteur (répétitions, pauses silencieuses... voir Siegman 1979; Guaïtella 1991), seule l'hésitation sonore a un statut d'exclusivité. En effet, elle seule ne peut signifier autre chose que le témoignage perceptible de l'activité de planification du locuteur (ou de l'utilisation codée et métaphorique de cette activité; par exemple lorsque quelqu'un a préparé son texte à l'avance et qu'il le dit en ajoutant des hésitations sonore pour le rendre plus "naturel"). Si l'hésitation vocale n'a pas de statut autre, elle se double cependant de la fonction de maintien de la parole pour le locuteur pendant la planification (Maclay, Osgood 1959). Un des points sur lesquels nous avons centré notre étude de l'intonation en parole spontanée, est l'observation de la réalisation vocale des hésitations. Nous différencions l'hésitation vocale, non seulement de la pause silencieuse, mais aussi de l'ensemble des phénomènes sonores présents dans la parole - tels que bruits de bouche divers, insertion d'éléments vocaliques non justifiés par le niveau segmental ... - qui nécessiteraient également d'être étudiés, et qui, à notre avis, n'ont pas (ou du moins pas seulement) fonction d'hésitation. Parmi ces

derniers, nous classons également les "heu" brefs que nous n'étudierons donc pas ici (voir conclusion). Nous pensons que la manifestation vocale de l'hésitation joue un rôle spécifique dans l'organisation rythmique de la communication. La caractéristique de ce phénomène est de se produire sous la forme d'une variation linéaire, descendante, et de faible amplitude de la fréquence fondamentale, parvenant parfois jusqu'à une instabilité périodique que l'on peut assimiler à de la "creaky voice". Notre hypothèse est que cette "intonation de l'hésitation" n'est autre qu'une manifestation de la ligne de déclinaison (dans le sens purement physiologique de ce terme), et que ceci pourrait expliquer les fonctions et la perception du phénomène d'hésitation dans la parole. Nous proposerons, en conclusion, une interprétation des phénomènes vocaliques de durée très brève qui sont parfois assimilés à des hésitations, et que nous n'étudions pas ici. Remarquons qu'actuellement les besoins en reconnaissance automatique de la parole nécessitent d'affronter la réalité de l'hésitation vocale et d'en aborder la description acoustique (Carbonnel 1991). Nous avons montré (Guaïtella 1989, 1991) que l'étude de la prosodie en parole spontanée requérait des grilles d'analyse établies d'après un modèle acoustico-perceptif. Ce point de vue, qui coïncide avec les préoccupations liées au dialogue homme-machine, est également à l'origine de notre étude des hésitations vocales.

2- METHODOLOGIE: LA DETERMINATION DES INDICES

On entend souvent dire qu'il est difficile de distinguer, sur une représentation graphique du signal acoustique, une syllabe accentuée allongée d'une hésitation sonore. Cependant, ni le locuteur en situation de communication, ni le phonéticien lorsqu'il écoute le document sonore, ne confondent une syllabe accentuée et une hésitation. Il nous semble donc indispensable de déterminer des critères distinctifs pour les deux phénomènes. Nous pensons que la différence doit se manifester au niveau de la gestion des paramètres (durée segmentale et fréquence fondamentale) par le locuteur, et non pas au niveau d'une interprétation contextuelle. Ce qui autorise donc une description objective paramétrisable. Nous avons considéré les paramètres de f0 et de durée segmentale, à travers leur participation à la production de syllabes accentuées et d'hésitations vocales. Nous pensons, même si nous ne les avons pas prises en compte, que les variations d'intensité pourraient également être étudiées dans cette perspective. Cette préoccupation de la détermination des faits à partir des indices correspond à l'idée qu'il n'existe pas de structure sous-jacente permettant de déterminer l'organisation rythmique en parole spontanée. Celle-ci ne peut qu'être observée a posteriori en fonction de grilles permettant de gérer les indices observables.

2-1- LA FREQUENCE FONDAMENTALE

Notre point de vue étant que la localisation de l'accent n'est pas prédictible par des règles issues de composantes verbales (Bolinger 1972), le problème est de savoir comment situer les syllabes mises en relief. La détermination de l'organisation rythmique *à partir des indices* devient alors un passage obligatoire sans lequel nous ne saurions parler d'analyse rythmique, et non l'occasion d'une vérification à posteriori de l'activité des paramètres en fonction d'une structure rythmique prédéterminée. Le rythme suppose une série d'éléments discrets entretenant des relations réciproques. Il va donc être nécessaire de retrouver cette série d'éléments discrets à l'intérieur d'un mouvement continu. Si une courbe se réalise sur plusieurs syllabes, nous pouvons considérer que ces syllabes constituent la suite d'éléments en question. Nous pourrions également envisager de découper le continuum temporel en segments de durée égale, ou se baser sur des critères morpho-syntaxiques, etc... La théorie de l'information propose des outils conceptuels pour répondre à ce type de question. Si nous considérons une courbe, sans même chercher à la segmenter en unités de taille inférieure (syllabes ou segments temporels), nous pouvons voir, grâce à l'analyse en termes "d'information", que cette courbe dispose d'une dynamique rythmique interne. Imaginons un continuum de fréquence fondamentale. Nous pouvons constater que l'apport d'information est constitué par le démarrage, l'arrêt ou le changement de sens d'une courbe et par le niveau fréquentiel ou ceux-ci se réalisent. Nous considérons que ces changements dans la trajectoire participent en tant qu'indices à l'organisation rythmique. Leurs positions par rapport à la chaîne segmentale permettra de déterminer les syllabes mises en relief. Par rapport au modèle d'interprétation proposé par la théorie de l'information, notre méthodologie présente des lacunes qu'il sera nécessaire de combler. Parmi celles-ci, la notion de probabilité d'apparition d'un phénomène. Or, cette notion est très délicate à manipuler en prosodie, il s'agit en effet de déterminer ce qui est prévisible et ce qui ne l'est pas dans l'évolution de la fréquence fondamentale. Etant donné que la fréquence fondamentale est directement contrainte par la production physique de la voix, on peut supposer que tout ce qui, dans la fréquence fondamentale, est conditionné physiologiquement, apparaît comme prévisible. Par exemple, le fait que l'intonation ne puisse pas monter indéfiniment (...) est conditionné physiologiquement, et est, par conséquent, prévisible pour le locuteur/auditeur possédant une connaissance empirique de ce phénomène. Autre exemple: pour produire deux montées successives, il est bien entendu nécessaire de produire une descente (ou un silence), ceci est également prédictible puisque le locuteur a une connaissance à-priori de ce phénomène. Or, pour l'ensemble de ces phénomènes, il est difficile d'évaluer la limite de ce qui est contraint, et donc prédictible, et de ce qui fait l'objet d'un choix expressif,

et donc est réellement porteur d'information. Ce problème nous semble pourtant essentiel, puisqu'il nous amène à appréhender les phénomènes intonatifs en tant que trace d'un "conflit" entre volonté expressive et contrainte physiologique, la première étant non-prédictible, la seconde prédictible, ce conflit s'incarnant dans l'évolution du système, perturbé par les conditions externes mais inéluctables, qui régissent son fonctionnement. Ainsi, les facteurs externes (situation de parole, émotion...) agissent sur le système prosodique. Sans ces facteurs, la parole perdrait son caractère "informatif". Peut-on et doit-on décrire le continuum vocal en termes de segments, ce qui permettrait de déterminer ceux qui sont porteurs d'information? La segmentation en syllabes permet de répondre objectivement, et, au moins, provisoirement, à ce type de question. La nature d'unité perceptive de la syllabe a été attestée par plusieurs études (Segui, Frauenfelder, Mehler 1981; Santi 1987; Santi, Cavé 1988).

2-2- LA DUREE

La modification de durée syllabique est attestée par de nombreuses études comme étant un phénomène essentiel de la manifestation du rythme. Or, notre travail sur la durée de syllabes accentuées en lecture de phrases (Guaïtella 1988, 1991), nous amène à considérer que les modifications de durée des syllabes accentuées ne consistent pas seulement en un allongement mais parfois en un raccourcissement. Comment, donc, déterminer la présence d'un accent à partir des indices fournis par la durée syllabique, si celle-ci peut varier à la fois dans le sens de l'allongement et du raccourcissement? Nous avons tenté de répondre à cette question à partir des deux principes suivants:

- la pause silencieuse apparaît comme une remise à zéro de la valeur syllabique de référence. Après chaque pause silencieuse il sera donc nécessaire de reconsidérer le rapport entre durée syllabique de référence (durées des syllabes atones) et durées particulières (durées des syllabes accentuées). Ce principe nous permet de faire face aux problèmes de variations globales de la durée syllabique, engendrant des modifications dans le pourcentage de variation de durée entre syllabes atones et syllabes accentuées. Cette procédure se justifie par le fait que la pause silencieuse, par sa durée propre, permet l'effacement dans la mémoire à court terme de la trace de la durée syllabique de référence correspondant à ce qui précède la pause. En outre, nous pensons que la pause est nécessaire pour que le locuteur lui-même parvienne à établir une nouvelle valeur de référence.

- Suite à une pause silencieuse, nous considérons donc le pourcentage de variations entre la première et la deuxième syllabe. Si ce pourcentage est inférieur à 20% (Rossi 1972; Klatt 1976), nous considérons alors la variation entre la deuxième et la troisième syllabe, et ainsi de suite jusqu'à ce que nous disposions d'une variation supérieure à 20%. Dans ce cas, nous considérons qu'il s'agit de l'établissement du "pourcentage de variation référentielle", et nous appliquons ensuite ce pourcentage pour chaque syllabe par rapport à la syllabe qui précède, jusqu'à la pause silencieuse suivante, afin de déterminer si la durée syllabique est un indice accentuel ou non.

Exemple: (extrait du corpus *Madagascar*, Guaïtella 1991)
segmentation syllabique avec durées des syllabes et indices de durée (représenté par [+])
"#
[c'est-à] - 219 ms
[dire] - 212 ms (la différence de durée avec la syllabe précédente est inférieure à 20%)
[que] - 140 ms (la différence de durée entre cette syllabe (s2) et la syllabe précédente (s1) est supérieure à 20%. Cette différence va permettre d'établir la valeur de référence pour comparer les durées syllabiques, la valeur de référence sera égale dans ce cas à s1/s2=1.5) [+]
[les] - 114
[mal] - 204 [+]
[gaches] - 229
[sont] -168
[cons] - 180
[tem] - 147
[ment] - 211 [+]
[tour] - 140 [+]
[nés] - 123
[vers] - 127
[la] - 140
[france] - 495 [+]
#"

L'ensemble de cette procédure - la référence à la pause qui précède et aux durées syllabiques initiales - est régie par le principe de sélection à gauche des composantes de référence, c'est-à-dire que tout est organisé en fonction de ce qui précède. Ceci est donc applicable, selon nous, à la parole spontanée. Le principe de sélection à droite - et la planification de la phrase - peut modifier complètement le problème en lecture (sélection à droite car tout est organisé en fonction de ce qui suit et qui est prévu à l'avance), puisque le locuteur sait alors qu'il va produire, par exemple, une syllabe accentuée en fin de phrase. On peut supposer qu'il programme dès le début de la phrase la variation référentielle de durée, mais qu'il réserve l'application de l'indice à la syllabe qu'il veut accentuer. Notre hypothèse est que lorsque l'indice de durée segmentale est utilisé avec l'indice de f0, il ne peut s'agir que d'une syllabe accentuée, alors que lorsque l'indice de durée est utilisé seul, il peut s'agir d'une hésitation vocale. C'est donc à travers l'étude de l'ensemble de l'organisation rythmique que l'on peut dégager les spécificités d'un élément fonctionnel.

3- APPLICATION: ÉTUDE D'UN EXTRAIT D'INTERVIEW

Cette étude porte sur un extrait d'interview, enregistrée en chambre anéchoïque, d'une locutrice francophone dépourvue d'accent régional marqué. Il s'agit d'un récit de voyage. Nous n'envisageons pas ici une

description rythmique globale (voir Guaïtella 1991), mais uniquement l'observation des hésitations vocales.

3-1- CONFIGURATIONS DE FRÉQUENCE FONDAMENTALE ET CONTEXTES D'APPARITION DE L'HÉSITATION

Nous avons observé la totalité des hésitations en fonction de:

- leur contexte d'apparition et de disparition (précédée ou suivie de texte ou de silence)
- leur durée
- la configuration et l'évolution de leur fréquence fondamentale.

L'étude systématique porte donc sur une seule locutrice, cependant nous avons validé les tendances générales de nos résultats par l'observation de données complémentaires issues de deux corpus d'interview, réalisées pour la télévision, de deux locuteurs méridionaux. Les occurrences relevées nous ont permis de déterminer que les hésitations peuvent intervenir dans tous les contextes, c'est-à-dire:

- précédées et suivies de silence (22.5% des cas)
- précédées de parole et suivies de silence (50%)
- précédées de silence et suivies de parole (5%)
- précédées et suivies de parole (22.5%)

L'évolution de la fréquence fondamentale nous a permis de dégager quatre types de configurations:

- la simple déclinaison de f0 (62.5% des cas)
- la déclinaison se terminant par de la creaky voice (17.5%)
- la déclinaison "modulée", c'est-à-dire avec de très petites montées de f0, suivies de déclinaisons plus importantes et de durée significative (12.5%) -
- la déclinaison se terminant par une "amorce", c'est-à-dire une montée relativement importante de f0 (7.5%)

Pour les deux derniers types de configurations, la durée dépasse déjà celle d'un allongement syllabique significatif, avant qu'intervienne une remontée de f0. Pour le dernier cas, nous considérons qu'un même segment vocalique se répartit en deux composantes: une hésitation puis une syllabe accentuée. Ce point de vue se justifie par l'impression auditive et par la "logique linguistique" du phénomène. On imagine aisément, en effet, que le locuteur puisse hésiter sur un segment vocalique, puis accentuer ce même segment pour initialiser ce qu'il va dire.

3-2- HÉSITATION ET VARIATION DE DURÉE

L'observation de la durée des hésitations montre que celles-ci sont d'une durée plus grande que celle des (autres?) syllabes (tableaux 1 et 2). La durée minimale relevée étant de 197 ms, et la durée maximale de 1157 ms.

Légende:

T : présence de texte
: présence d'une pause
- # : hésitation comprise entre deux pauses
\ : décroissance progressive de la fréquence fondamentale
k : présence de creaky voice
m : présence de modulations de f0 sur la décroissance
Am : présence d'une amorce de montée de f0 en fin de décroissance

	moy	min	max
T - #	465.4	197	712
# - T	574	420	728
T - T	413.9	237	579
# - #	657.7	393	1157

tab. 1: Durée des hésitations vocales en fonction des contextes d'apparition (en ms).

	moy	min	max
\	476.4	197	930
k	567.2	237	1157
m	527	420	712
Am	582.7	488	756

tab. 2: Durée des hésitations vocales en fonction des configurations prosodiques (en ms).

Ce qui ressort de l'observation de ces valeurs, est essentiellement la durée plus importante des hésitations comprises entre deux pauses silencieuses. La durée des hésitations (en moyenne sur l'ensemble des valeurs observées: 533 ms) apparaît comme intermédiaire entre la durée syllabique (moyenne des durées syllabiques pour le corpus: 184.5 ms) et la durée des pauses silencieuses (durée moyenne des pauses silencieuses pour le corpus: 817.1 ms). Il semble que l'on puisse décrire ce phénomène comme la coexistence de trois zones de durées segmentales différentes: la zone temporelle des durées syllabiques, celle des durées d'hésitation, et celle des durées de pauses, l'ensemble de ces durées catégorielles étant toujours relativisables contextuellement.

3-3- HÉSITATION ET VARIATION DE FRÉQUENCE FONDAMENTALE

La fréquence fondamentale chute toujours sur les hésitations (quelques rares cas de contours totalement plats, mais jamais de montée), cependant, l'importance de cette chute peut varier (tableaux 3 et 4).

	moy	min	max
T - #	- 28.6	- 10	- 67
# - T	- 48.5	- 42	- 55
T - T	- 22.75	0	- 50
# - #	- 45	- 25	- 65

tab. 3: Variations de la courbe de f0 en fonction du contexte d'apparition (en Hz).

	moy	min	max
\	- 30.75	0	- 65
k	- 32.65	- 10	- 50
m	- 34.8	- 20	- 67
Am	- 39	- 27	- 60

tab. 4: Variations de la courbe de f0 en fonction des configurations prosodiques (en Hz) - *pour les "amorces", nous avons relevé la dernière valeur avant le remontée de f0, et pour les "modulations", nous avons considéré la décroissance globale de f0.*

Lorsque l'hésitation est précédée et suivie de texte, elle présente une déclinaison moins importante (voire nulle), de même, cette dernière est également plus brève, ce qui n'est pas surprenant. L'hésitation précédée de pause (suivie de texte ou de pause) présente la déclinaison la plus forte et la durée la plus longue. Il semble donc que l'importance de la déclinaison soit liée à ce qui encadre l'hésitation, indépendamment de la configuration prosodique réalisée.

4- PERSPECTIVES: CONSÉQUENCES SUR LA PERCEPTION DU RYTHME

L'hésitation correspond à la volonté de conserver l'"espace sonore" pendant que l'on élabore la suite de son propos (Maclay, Osgood 1959). La particularité de l'hésitation est de présenter une émission vocale non-verbale, nécessaire pour maintenir le contact, tout en "gardant la parole", pendant une période de temps que le locuteur réserve uniquement à la conceptualisation. La production vocale, pendant ce temps, s'assimile à une tenue - c'est d'ailleurs l'impression que l'on a à l'audition des séquences d'hésitation - or, si cette "tenue" volontaire se présente, au niveau acoustique, sous la forme d'une déclinaison de f0, c'est que cela correspond à une diminution de la pression sous-glottique, c'est-à-dire à la dimension physiologique de la ligne de déclinaison. Nous nous référons ici au phénomène décrit par Lieberman (voir notamment Lieberman et al. 1985), c'est-à-dire l'organisation de la ligne de f0 en relation avec le groupe de souffle, et non pas à la programmation d'unités textuelles en relation avec f0. Les "heu brefs" pourraient être interprétés comme des hésitations manifestées par un indice de durée de type raccourcissement (Guaïtella 1988, 1991). Cependant cette interprétation nous semble peu probable, le propre de l'hésitation étant, en effet, de fournir le temps nécessaire à la planification du discours (se manifestant pendant les "P-phases" selon la terminologie de Butterworth et Goldman-Eisler, 1979). Nous interprétons plutôt ces cas comme l'insertion d'éléments vocaliques dont la fonction ne serait pas l'hésitation, et que l'on pourrait interpréter soit comme des mises en valeur lexicales, soit comme des ajouts dans un but d'équilibrage rythmique... Ils pourraient être rapprochés des phénomènes du type claquements de langue, bruits de bouche... La réalisation acoustique de l'hésitation vocale nous semble intéressante, car il s'agirait d'une émanation d'une activité exclusivement physiologique dans la procédure de parole. Ceci pourrait expliquer pourquoi les hésitations sont difficilement mémorisables - si ce n'est difficilement perceptibles - dans une situation de communication normale. Il est bien connu qu'on ne prête attention aux hésitations que si elles sont extrêmement fréquentes. Chez un locuteur qui produit des hésitations selon une fréquence habituelle, on ne remarque pas celles-ci (Blanche-Benveniste 1985), et lui-même n'en est pas conscient. En outre, lorsque l'on transcrit de la parole spontanée, il est nécessaire de repasser plusieurs fois l'enregistrement et de se concentrer particulièrement pour repérer, localiser et compter les hésitations. Or, si on admet que la manifestation vocale de l'hésitation est de nature exclusivement physiologique, cela permettrait d'interpréter l'absence de mémorisation des hésitations comme l'exclusion du phénomène en dehors du niveau segmental (niveau verbal du continuum syllabique), ce qui ne signifie pas que ce dernier soit sans fonction communicative. En outre, le fait que l'hésitation se situe à l'intérieur d'une zone de durée particulière, faciliterait un traitement perceptif spécifique. Cette interprétation n'est pas surprenante si l'on songe au fait suivant: les hésitations ne sont pas - ou rarement - produites volontairement. On peut donc concevoir qu'il existe un accord inconscient entre les locuteurs en faveur de la non-mémorisation des hésitations, lesquelles sont inévitables et indispensables dans la communication, mais, également et inversement, vécues comme des "défaillances" dans le raisonnement. Leur rôle dans la communication (et plus spécialement dans la perception par l'interlocuteur) est probablement de fournir un "temps de repos" aisément repérable par ces caractéristiques acoustiques, pendant lequel l'interlocuteur (au même titre que le locuteur) pourra faire le point sur ce qui a été dit, et inférer sur ce qui va être dit.

5- CONCLUSIONS

Si on utilise la méthodologie, décrite ci-dessus, de détermination de la place de l'accent à partir des indices, on s'aperçoit que l'hésitation se caractérise par un indice significatif et spécifique d'allongement de la "syllabe", en l'absence d'indice fourni par la fréquence fondamentale (pas de "décrochage"). L'allongement de la durée syllabique sans décrochage de la fréquence fondamentale, ne serait donc pas un indice d'accentuation, mais - du moins pour les cas d'accroissement extrême - un indice d'une autre catégorie segmentale: l'hésitation. On peut se demander si l'établissement d'un seuil permettant (de façon locale et contextuelle) de différencier, d'une part, syllabes non-accentuées avec allongement de durée, et d'autre part, hésitation vocale, est envisageable. On observe deux cas possibles pour les hésitations:

- une durée segmentale démarcative sans décrochage de F0

- un décrochage sur le début ou la fin d'une syllabe, suivi ou précédé d'un prolongement significativement important de ce segment, sans décrochage. On peut alors considérer que la même unité segmentale se divise en deux catégories perceptives successives: une syllabe (accentuée) suivie ou précédée d'une hésitation.

Nous pouvons considérer qu'il existe (au moins) trois unités au niveau segmental - pause, hésitation, syllabe -, et que ceci serait confirmé par l'utilisation du paramètre de durée qui permet la différenciation de ces trois catégories du point de vue perceptif. Le paramètre de durée apparaît comme un indice accentuel, mais demeure uniquement un indice de second plan associé à la présence d'un décrochage de f0. L'existence du phénomène d'hésitation confirme le rôle primordial du paramètre de f0 pour la perception de l'accent.

BIBLIOGRAPHIE:

AUTESSERRE D., NISHINUMA Y., GUAITELLA I., 1989, "Breathing, pausing, and speaking in dialogue", Proceedings of the Eurospeech Conference, Paris.

BAILLY G., BENOIT C., (à paraître), *Talking Machines*, Elsevier North-Holland.

BLANCHE-BENVENISTE C., 1985, "La dénomination dans le français parlé: une interprétation pour les 'répétitions' et les 'hésitations'", *Recherches sur le français parlé*, 6, Groupe Aixois de Recherche en Syntaxe, Université de Provence.

BOLINGER D.L., 1972, "Accent is predictable (if you're a mind-reader)", *Language*, 48, 633-44.

BOLINGER D.L., 1985, *Intonation and its parts*, Edward Arnold.

BUTTERWORTH B., 1980, "Evidences from Pauses in Speech", in Butterworth *Language Production*, vol.1 *"Speech and Talk"*, Academic Press, 155-76.

BUTTERWORTH B., 1980, *Language Production*, vol.1 *"Speech and Talk"*, Academic Press, London.

BUTTERWORTH B., GOLDMAN-EISLER F., 1979, "Recent Studies in Cognitive Rhythm", in Siegman, Feldstein, *Of speech and time Temporal Speech Patterns in Interpersonal Contexts*, Lawrence Erlbaum Associates, Hillsdale, New Jersey, 211-24.

CARBONNEL N., 1991, "Détection d'informations prosodiques suprasegmentales pertinentes en reconnaissance-compréhension automatique de la parole continue", *Prépublication des Actes du Séminaire Prosodie du GRECO PRC-HM, Pole Parole*, Aix-en-Provence.

DAVIS D., LEON P., 1989, "Pausologie et production linguistique", *Information Communication*, 10, Toronto, 31-43.

FRAISSE P., 1974, *Psychologie du rythme*, P.U.F.

GOLDMAN-EISLER F., 1958, "Speech Analysis and Mental Processes", *Language and Speech*, 1, 59-75.

GOLDMAN-EISLER F., 1968, *Psycholinguistics Experiments in spontaneous speech*, Academic Press.

GOLDMAN-EISLER F., 1972, "Pauses, Clauses, Sentences", *Language and Speech*, 15, 103-13.

GUAITELLA I., 1988, "Variation de durée en syllabe accentuée" *Travaux de l'Institut de Phonétique d'Aix*, 12, 185-204.

GUAITELLA I., 1990, "Propositions pour une méthode d'analyse de la fréquence fondamentale en parole spontanée", *Actes du premier congrès français d'acoustique*, Lyon, Les éditions de Physique, 515-8.

GUAITELLA I., 1990, "Composants rythmiques para-verbaux dans la parole", *Proceedings of the LP'90 Conference*, Prague (actes à paraître).

GUAITELLA I., 1991, *Rythme et parole: comparaison critique du rythme de la lecture oralisée et de la parole spontanée*, Thèse, Aix-en-Provence.

GUAITELLA I., AUTESSERRE D., NISHINUMA Y., 1990, "Some physiological aspects of pausing and speaking in dialog", *J. Acoust. Soc. Am.*, Sup.1, vol. 87.

GUAITELLA I., SANTI S., 1990, "Ponctuation et organisation rythmique de l'oral", *Proceedings of the LP'90 Conference*, Prague (à paraître).

GUAITELLA I., SANTI S., (à paraître), "The punctuation and perception of read and spontaneous prosody: an application to speech synthesis", in: Bailly, Benoit, *Talking Machines*, Elsevier North-Holland.

KLATT D., 1976, "Linguistic uses of segmental duration in English: acoustic and perceptual evidence", *J. Acoust. Soc. Am.*, 59 (5), 1208-21.

LIEBERMAN P., KATZ W., JONGMAN A., ZIMMERMAN R., MILLER M., 1985, "Measures of the sentence intonation of read and spontaneous speech in American English", *J. Acoust. Soc. Am.*, 77 (2), 649-57.

MACLAY H., OSGOOD C.E., 1959, "Hesitation phenomena in spontaneous English speech", *Word*, 15, 19-44.

MOLES A. A., 1967, "Informatique du rythme", *Actes du colloque sur Les Rythmes*, Lyon, 275-89.

ROSSI M., 1972, "La perception de la durée et ses implications phonétiques", *Travaux de l'Institut de Phonétique d'Aix* , 1.

SANTI S., 1987, *La syllabe, unité perceptive: étude expérimentale*, Mémoire de D.E.A., Université de Provence.

SANTI S., CAVE C., 1988, "Segmentation syllabique et niveau de réalisation de la tâche", *Travaux de l'Institut de Phonétique d'Aix*, 12, 105-14.

SANTI S., GUAITELLA I., 1990, "Variations of duration in stressed syllables taken from French read sentences", *J. Acoust. Soc. Am.*, Sup. 1, vol. 87.

SEGUI J., FRAUENFELDER U., MEHLER J., 1981, "Phoneme monitoring, Syllable monitoring and Lexical access", *British Journal of Psychology* .

SHANNON R., WEAVER W., 1949, *The Mathematical Theory of Communication*, Illinois University Press.

SIEGMAN A.W., 1979, "Cognition and Hesitation in Speech", in: Siegman, Feldstein, *Of Speech and Time*, Lawrence Erlbaum Associates, Hillsdale, New Jersey, 151-78.

SIEGMAN A.W., FELDSTEIN S., 1979, *Of Speech and Time*, Lawrence Erlbaum Associates, Hillsdale, New Jersey.

19es J.E.P. - Bruxelles, 19 au 22 mai 1992.

MODÉLISATION, PAR UN SYSTEME DYNAMIQUE, DE TRAJECTOIRES ACOUSTIQUES UNIDIMENSIONNELLES

M. PITERMANN*, J. CAELEN**

*INSTITUT DE PHONÉTIQUE, UNIVERSITÉ LIBRE DE BRUXELLES
**INSTITUT DE LA COMMUNICATION PARLÉE, GRENOBLE

Résumé

La variabilité contextuelle et interlocuteurs constitue un obstacle à la réalisation d'un décodeur acoustico-phonétique. Transposer la notion de cible articulatoire au niveau acoustique pourrait concourir à la solution de ce problème [Caelen, 1985]. Malheureusement, l'extraction des pseudocibles acoustiques n'est pas encore satisfaisante. Modéliser les trajectoires d'indices par un système dynamique devrait permettre une identification des pseudocibles acoustiques comme points limites dans un espace de représentation adéquat. En effet, la comparaison entre l'évolution temporelle d'un indice acoustique et celle de la solution asymptotique locale d'un modèle linéaire suggère que l'indice évolue de pseudocible en pseudocible sous l'influence d'un système dynamique linéaire non stationnaire.

Introduction

En reconnaissance automatique de la parole, le signal est trop riche en informations pour pouvoir être traité tel quel. Il doit donc être prétraité. Généralement on le réduit à un jeu d'indices (formants, indices spectraux, articulatoires...) contenant le maximum d'informations pertinentes pour l'application en éliminant au mieux le superflu (variabilité contextuelle, interlocuteurs...). Ensuite on recherche des structures invariantes dans les indices afin de dégager des règles.

J. Caelen a émis l'hypothèse que la notion de cible articulatoire serait transposable au niveau acoustique [Caelen, 1985], nous parlerons alors de pseudocible. Ces pseudocibles sont définies comme les valeurs vers lesquelles les indices acoustiques tendraient lors de la réalisation d'un segment acoustique. En DAP, l'extraction des pseudocibles permettrait peut-être d'apporter une solution au problème de la variabilité contextuelle et interlocuteurs.

Trois problèmes se posent dans cette approche:
i) Tous les segments acoustiques ne donnent pas lieu à un comportement asymptotique des indices.
ii) Le principal moyen d'extraction des pseudocibles, c'est-à-dire la recherche des extrema locaux et des points d'inflexions des trajectoires d'indices, est un problème ardu dans le cas de signaux bruités.
iii) Les indices sont parfois loin d'atteindre leurs pseudocibles, et leur détermination dans ce cas est sans solution à ce jour.

Pour résoudre les points (ii) et (iii), nous partons d'une idée suggérée par l'approche de J. Caelen : un système dynamique non stationnaire pourrait gouverner l'évolution temporelle des indices. Dans ce cas, si pour un segment acoustique le système dynamique reste stationnaire pendant qu'un indice s'approche de sa pseudocible, ce segment serait caractérisable par la solution asymptotique du système dynamique.

Notre but est d'étudier la question suivante : "Y a-t-il un système dynamique, stationnaire pour certains segments, qui modélise l'évolution temporelle d'un indice acoustique ?". Dans cet article nous nous intéressons au cas unidimensionnel.

Le système dynamique est exprimé par un système d'équations différentielles : $\dot{x} = f(x,t)$ où x(t) est un vecteur qui décrit l'état du système à l'instant t. Lorsque les données sont discrètes, les équations différentielles sont remplacées par des équations aux différences. En cas de stationnarité du système, on appelle solution asymptotique un état stationnaire vers lequel il tend asymptotiquement lorsque le temps tend vers l'infini. Les systèmes les plus simples qui possèdent ces solutions sont les modèles linéaires avec forçage constant:

$$y_n = a0 + \sum_{i=1}^{ordre} a_i y_{n-i} \qquad (1)$$

Le forçage a0 est nécessaire pour pouvoir modéliser une autre solution asymptotique x* que 0:

$$x^* = \frac{a0}{1 - \sum_{i=1}^{ordre} a_i} \qquad (2)$$

Les résultats préliminaires montrent que le second ordre suffit quand on s'intéresse aux asymptotes très approchées.

Le signal de parole est un processus non stationnaire. Dès lors, les paramètres de notre modèle vont évoluer au cours du temps. Dans ce cas, on ne peut plus parler rigoureusement de solution asymptotique du modèle, mais on peut toujours, pour une fenêtre d'analyse donnée, estimer les paramètres du modèle, et obtenir ainsi une estimation locale de la solution asymptotique.

En estimant x* pour une série de fenêtres temporelles que l'on déplace d'un échantillon à la fois, on peut construire une série chronologique x*(t). Pour un signal stationnaire, les estimations de x* sont à peu près identiques, et x*(t) prend localement la forme d'un plateau, même si l'indice est loin de sa pseudocible Dans le cas contraire, une solution asymptotique n'existe pas, et x*(t) peut prendre localement n'importe quelle forme. En d'autres termes, si un modèle linéaire avec forçage convient pour modéliser l'évolution temporelle d'un indice et si ce modèle est stationnaire, x*(t) devrait présenter un palier.

x*(t) permettrait donc:

i) D'indiquer les segments acoustiques qui satisfont nos hypothèses.

ii) De remplacer la recherche d'extrema locaux et de points d'inflexions par celle de paliers qui sont plus simples à détecter et qui offrent l'avantage de fournir les valeurs des asymptotes et les frontières des segments qui y sont associés.

iii) De trouver peut-être des pseudocibles non atteintes.

iv) D'effectuer une réduction de données sur l'indice.

Nous nous intéressons à un ensemble de sept indices acoustiques (Energie, A/G, F/O, B/D, E/C, D/S, C/D) [Caelen et al., 81] définis statistiquement par des combinaisons linéaires d'énergies calculées à partir d'un modèle d'oreille à 24 bandes [Caelen, 79]. Ces indices offrent l'avantage de condenser l'information pertinente au prix de la perte d'une réelle interprétation phonétique.

Dans cet article, nous justifions le choix du modèle grâce aux résultats obtenus par simulations sur ordinateur. Ensuite nous présentons les résultats d'une modélisation de F/O sur un corpus phonétiquement équilibré de quelques phrases prononcées par un locuteur masculin. Les résultats suggèrent que l'indice évolue de pseudocible en pseudocible sous l'influence d'un système linéaire.

Méthode

Pour estimer localement les paramètres du modèle à partir des valeurs d'un indice, nous utilisons une régression multiple, c'est-à-dire une minimisation, au sens des moindres carrés, de l'erreur qui sépare les

données de la série chronologique générée par le modèle.

Afin d'étudier la méthode, nous avons synthétisé des séries temporelles bruitées à partir de modèles linéaires de différents ordres. Ensuite, nous avons utilisé des fenêtres d'analyse de différentes longueurs sur ces séries pour estimer la solution asymptotique locale x*.

Nous avons aussi testé l'approche avec l'indice F/O calculé à une fréquence d'échantillonnage de 125 Hz sur un corpus phonétiquement équilibré de quelques phrases prononcées par un locuteur masculin. Nous avons utilisé le modèle linéaire du second ordre avec forçage pour construire la série temporelle x*(t) que nous avons comparée à celle de l'indice. Nous avons utilisé pour cela une fenêtre d'analyse glissante contenant 5 échantillons, ce qui correspond au minimum possible pour une estimation analytique. En effet, il nous faut 3 échantillons consécutifs pour former une équation $y_n=a_0+a_1y_{n-1}+a_2y_{n-2}$. En utilisant 3 valeurs successives de n, nous obtenons les 3 équations indispensables à la détermination de a_0, a_1, et a_2, ce qui nécessite 5 échantillons. Nous avons aussi examiné l'évolution des résultats en fonction de l'allongement de la fenêtre d'analyse.

En déplaçant d'un échantillon à la fois une fenêtre d'analyse de longueur fixe sur l'indice F/O, on observe que la série chronologique x*(t) contient une succession de plateaux accompagnés à leurs extrémités de points aberrants isolés ou en couples. Leur présence semble due au manque de respect de l'hypothèse de stationnarité locale du modèle, c'est-à-dire à une rupture du modèle. Dans un premier temps, nous avons décidé de les éliminer par un filtrage médian à 5 points.

L'étape suivante consiste à extraire les paliers de l'évolution temporelle de x*. Pour cette tâche nous considérons qu'un échantillon de la série chronologique appartient à un palier si l'un des deux points qui suivent l'échantillon ou le précèdent en est distant de moins de 2 % de la pleine échelle de l'indice. Si deux échantillons appartenant au même palier sont séparés par un troisième échantillon plus éloigné, nous considérons, pour des raisons de continuité, que ce dernier fait partie du même plateau. Pour déterminer la valeur d'un palier, nous prenons la médiane des échantillons.

Résultats

Les simulations montrent que l'estimation de la solution asymptotique du modèle linéaire du second ordre avec forçage est robuste au bruit si l'indice s'approche de sa pseudocible à une distance inférieure à l'amplitude du bruit. Dans ce cas, l'estimation est très bonne quel que soit l'ordre du modèle qui a servi a simuler la série chronologique de l'indice. Lorsqu'au contraire la fenêtre d'analyse intercepte des valeurs de l'indice qui sont loin de la pseudocible, d'une part, la robustesse de x* au bruit diminue, d'autre part, il devient important que l'ordre du modèle d'analyse corresponde à celui du modèle de simulation. En effet, dans ce dernier cas, l'estimation de x* est légèrement biaisée en direction des valeurs de l'indice.

On peut aussi noter que les estimations des paramètres a_0, a_1 et a_2 sont beaucoup plus sensibles que x* au bruit et à la correspondance entre l'ordre du modèle d'analyse et de simulation.

On peut donc conclure des simulations que:
i) Pour l'estimation d'une pseudocible très approchée, l'ordre du modèle d'analyse n'est pas crucial,
ii) Dans le cas contraire, l'ordre du modèle d'analyse est important, et une analyse des résidus des régressions multiples s'impose.

L'estimation de x*(t) réalisée à partir de l'indice acoustique F/O est présentée à la figure 1.a. Le temps y est gradué en secondes pour la phrase "Il se garantira du froid avec un bon capuchon.". La figure 1.b montre la même série temporelle filtrée par le

médian à 5 points servant à éliminer les points aberrants.

La figure 2 présente en trait plein la première moitié de la série chronologique filtrée de la figure 1.b, l'échelle temporelle ayant été dilatée d'un facteur 2. L'ordonnée a aussi été dilatée. On comparera cette série à celle de l'indice représentée en pointillés. La segmentation phonétique réalisée par H. Tattegrain sur base du signal de parole et des indices acoustiques [Tattegrain, 90] a été juxtaposée.

La figure 3 montre les paliers extraits de x*(t) représentés en traits pleins comparés avec la série chronologique de l'indice représentée en pointillés pour la même phrase. Une frontière isolée d'un palier est indiquée par un trait vertical, tandis que deux frontières adjacentes sont reliées par un trait oblique. Paliers et segmentation phonétique se correspondent.

L'allongement de la fenêtre d'analyse n'a guère amélioré les résultats. En effet, l'estimation analytique utilise 5 échantillons, comme nous en calculons un toutes les 8 ms, une telle fenêtre d'analyse correspond déjà à une durée de 40 ms, ce qui excède celle de certains segments phonétiques. Ainsi en allongeant la fenêtre d'analyse, nous obtenons une série chronologique x*(t) plus lisse. Par conséquent, les paliers se déforment en faisant tendre la série chronologique x*(t) vers celle de l'indice.

Discussion

L'observation de paliers suggère que l'indice acoustique évolue de pseudocible en pseudocible sous l'influence d'un système dynamique linéaire non stationnaire. Nous n'avons pas encore évalué quantitativement la correspondance entre les paliers et les segments acoustiques car nous ne possédons actuellement qu'une segmentation phonétique.

Nous devons aussi mieux valider le modèle. En effet, les simulations ont montré que la modélisation telle que nous l'avons réalisée place facilement des paliers sur les extrema et les points d'inflexions d'une série temporelle. Dès lors, une analyse des résidus des régressions multiples s'impose afin de déterminer l'ordre correct du modèle, ce qui le validerait en même temps. Cette étape est également nécessaire pour aborder le problème des pseudocibles non atteintes car leurs positions ne sont estimées correctement que si le modèle correspond aux données.

Il convient d'émettre une deuxième réserve : la fréquence d'échantillonnage de l'indice est trop faible pour que nous puissions disposer d'un nombre suffisant de données dans les fenêtres d'analyse pour l'ensemble des segments acoustiques du corpus. Nous devons, en effet, mieux nous affranchir du bruit. Malheureusement lorsque l'indice est calculé à une plus haute fréquence d'échantillonnage, il est entaché par des oscillations liées au signal glottique. Il est donc nécessaire dans ce cas de filtrer l'indice.

Mettre en oeuvre une méthode de détection de rupture du modèle nous semble aussi utile afin d'optimiser la longueur d'une fenêtre d'analyse de taille variable. En effet, à partir de la connaissance des intervalles de temps pendant lesquels le modèle est stationnaire, nous pourrions estimer x* à partir des plus longs segments de séries chronologiques possibles. Nous diminuerions ainsi l'influence du bruit sur les résultats, et nous éliminerions probablement les points aberrants observés aux frontières des paliers.

Conclusion

La modélisation permettrait:

i) De cataloguer les segments acoustiques qui satisfont nos hypothèses.

ii) De remplacer la recherche d'extrema locaux et de points d'inflexions par celle de paliers qui sont plus simples à détecter et qui offrent l'avantage de fournir les valeurs des asymptotes et les frontières des segments qui y sont associés.

iii) De trouver peut-être les pseudocibles non atteintes.

iv) D'effectuer une réduction de données sur l'indice.

Une première implantation de la méthode sur l'indice acoustique F/O a produit des résultats qui suggèrent que cet indice évolue de pseudocible en pseudocible sous l'influence d'un système dynamique linéaire non stationnaire.

Nous devons encore évaluer quantitativement la correspondance entre les paliers et les segments acoustiques. Nous devons aussi analyser les résidus des régressions multiples afin de trouver le bon modèle et le valider. Cette étape est aussi nécessaire pour traiter les pseudocibles non atteintes.

Une méthode de détection de rupture du modèle et une augmentation de la fréquence d'échantillonnage de l'indice s'avèrent nécessaires afin d'améliorer les résultats.

Références

[Caelen, 79] J. Caelen, *Un modèle d'oreille. Analyse de la parole continue. Reconnaissance phonémique.* Thèse d'Etat, Toulouse, 1979.

[Caelen et al, 81] J. Caelen et G. Caelen, *Indices et propriétés dans le projet ARIAL II.* Proceedings GALF-CNRS "Processus d'encodage et de décodage phonétiques", 1981, pp. 129-143, C. Abry, J. Caelen, J.S. Liennard, G. Perrenou, M. Rossi eds.

[Caelen, 85] J. Caelen, *Segmentation and Cinematic*, International Congress of Acoustic, Toronto, 1985.

[Tattegrain, 90] H. Tattegrain, *Un système expert de décodage acoustico-phonétique de la parole continue.* Thèse de Doctorat, pp. 81-82, Grenoble, 1990.

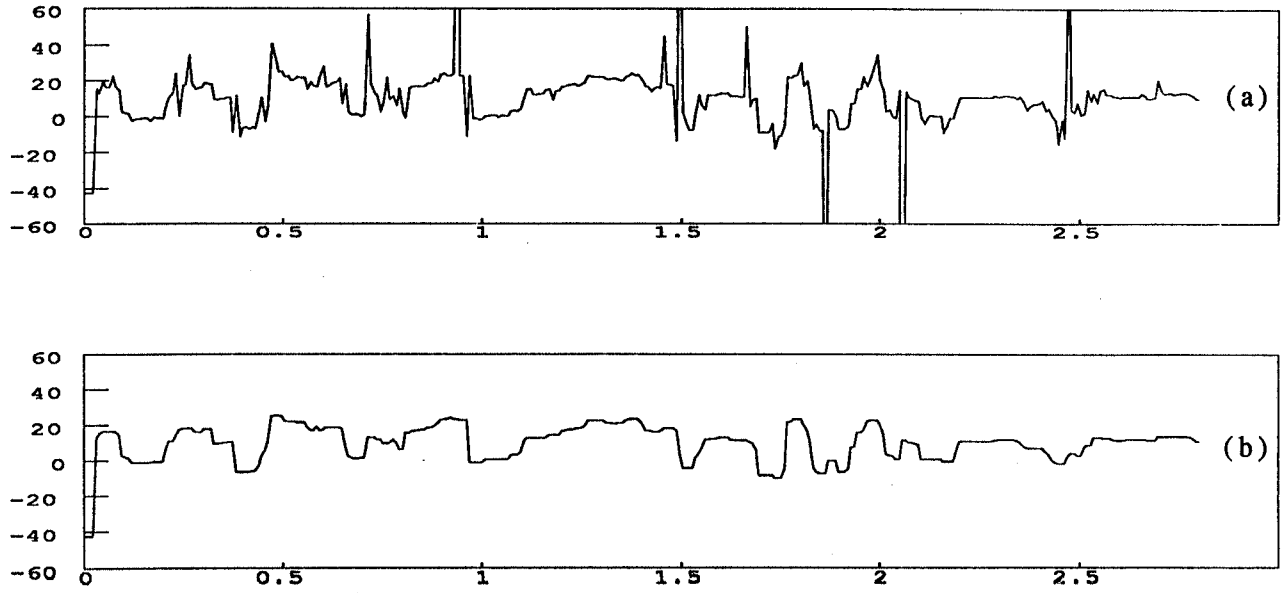


Figure 1. Comparaison entre la série temporelle de l'asymptote locale du modèle :
a) Non filtrée
b) Filtrée par un médian à 5 points.

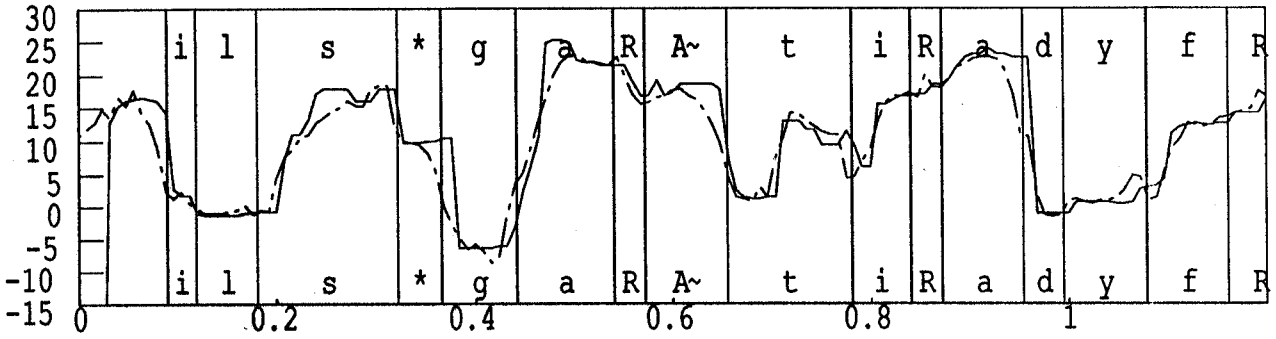


Figure 2. Comparaison entre la série temporelle de l'indice (pointillés) et celle de l'asymptote locale du modèle du second ordre avec forçage filtrée par un médian à 5 points (trait plein).

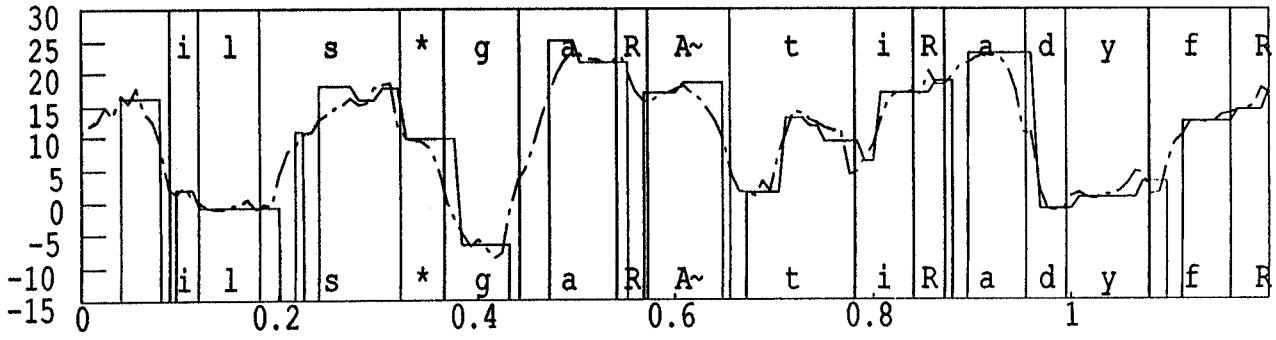


Figure 3. Comparaison entre la série temporelle de l'indice (pointillés) et les paliers extraits de l'évolution temporelle filtrée de l'asymptote locale du modèle du second ordre avec forçage (trait plein).

19es J.E.P. - Bruxelles, 19 au 22 mai 1992.

APPLICATION DE MODELES DE MARKOV A LA DESCRIPTION ARTICULATOIRE DU SIGNAL DE PAROLE

Ryszard GUBRYNOWICZ

Institut des Recherches Fondamentales de Technologie
Académie Polonaise des Sciences
rue Świętokrzyska 21, 00-049 VARSOVIE - Pologne

Résumé

On présente un système de description du signal de parole fondé sur une approche articulatoire. Come critère de classification de sons de la langue polonaise on a choisi en premier lieu le mode d'articulation, car ce trait est relativement peu sensible au phénomène de coarticulation. Pour l'apprentissage du système on a utilisé une liste de mots phonétiquement équilibrée. Pour ce corpus on a calculé les modèles de Markov pour 12 classes articulatoires. La vérification était réalisée par des expériences d'alignement effectuées sur un autre corpus de 20 phrases phonétiquement balancé. Les résultats préliminaires obtenus sont encourageants malgré la simplicité de modèles utilisés. Les avantages de la description articulatoire du signal de parole sont discutés.

1. INTRODUCTION

Au cours de dix dernières années on observe un interêt croissant à la modélisation des unités de parole de dimension sous-lexicale et leur application à la description du signal de parole représenté dans l'espace paramétrique. La raison de cette tendance est bien claire car l'opération avec un nombre limité d'unités permet de former des règles universelles de reconnaissance de la parole continue pour une langue donnée, indépendantes de dimensions du vocabulaire utilisé. Le problème du choix d'unités est une question toujours ouverte et cela malgré les succés évidents qu'on a obtenus avec les unités de dimension phonémique (par ex. [1,2]). Les difficultés viennent surtout du phénomène de la coarticulation qui pousse à augmenter le nombre d'unités de description au dessus du nombre de phonèmes, donc les mêmes phonèmes ont plusieurs réalisations de référence pour chaque contexte pris séparement.

Dépuis longtemps on est conscient que la notion de segmentation présuppose que l'onde de parole est constituée d'une suite de phonèmes, donc qu'elle est fondée sur un modèle extrêmement simpliste, non correspondant au conditions réels et ne tenant pas compte de tels phénomènes comme celui de coarticulation. C'est justement ce phénomène ne permet pas souvent de bien établir le début et la fin du segment d'onde sonore comme cela a lieu dans le cas de sons d'un caractère transitoire, et dans le cas extrême, même de les détecter quand il y a l'élision ou assimilation des consonnes. La raison capitale de cette complexité est la dichotomie entre le signal de parole qui est un résultat de l'activité du système articulatoire étant surtout une source acoustique et la description phonémique en unités abstraites qui est une forme du codage de l'information au niveau linguistique. C'est une de raison pour laquelle dépuis quelque temps, dans les systèmes de reconnaissance de parole on utilise un niveau intermediare de description, que'on peut appeler articulatoire.

Dans le communiqué on propose un système de description du signal de parole basé sur un nombre limité de classes articulatoires qui ne depasse pas une vingtaine. Cette approche permet d'abord d'écrire d'une façon robuste le signal de parole avant de passer à la reconnaissance de sons. Le mode d'articulation est choisi comme un trait principal de classification, qui est relativement peu sensible à la coarticulation. Cette approche était déjà

appliquée dans un système par règles [3,4,5]. Dans le travail actuel on a appliqué à la description articulatoire les modèles de Markov cachés calculés à l'aide de HTK Toolkit [6] et on a vérifié l'efficacité de cette approche par des expériences d'alignement.

2. LE CHOIX DU CORPUS ET PARAMETRISATION DU SIGNAL DE PAROLE

Chaque système qui est fondé sur la méthode HMM comporte deux phases, une d'apprentissage qui résulte en détermination des modèles statistiques pour chaque son (ou classe de sons) du language donné, et l'autre de reconnaissance, qui est en effet une vérification de modèles calculés. Ces modèles sont plus ou moins dépendants du locuteur. Contrairement aux autres pays, il n'existe pas en Pologne de base de données pour l'étude du language et reconnaissance de parole continue, accessible à tous les laboratoires et contenant un grand nombre de locuteurs. Pour cette raison on a decidé d'utiliser au départ des listes de mots pour lequelles les fréquences d'occurence de phonèmes sont proches de celles pour la langue naturelle. De telles listes sont utilisées dans les mesures de la nettété de la parole transmise par les lignes de télécommunication. En général, les mots dans ces listes sont prononcés en isolé. Une liste d'un tel type était utilisée pour l'apprentissage du système, mais les mots étaient prononcés d'une façon continue, en groupes par trois, en formant ainsi des pseudo-phrases. Le corpus d'apprentissage avait les mêmes fréquences d'occurence de phonèmes que pour la langue naturelle, mais les fréquences pour leur combinaisons en diades, triades et tétrades étaient bien différentes.

Le choix d'unités de description était orienté vers les unités le moins sensibles aux phénomènes de coarticulation. Pour cette raison on a choisi comme critère de classification de sons le mode d'articulation, de même que dans les travaux déjà mentionés [3,4,5]. Cette fois-ci on a choisi 12 classes articulatoires et 2 unités pour les silences, une précédante et subséquente la phrase, l'autre pour celle à l'intérieure de la phrase.

Le corpus d'apprentissage était enrégistré dans une chambre normale et calme, où le niveau de bruit était d'ordre de 40 dB plus bas par rapport au niveau moyen du signal. Les tests étaient réalisés avec d'autres corpus de phrases phonétiquement équilibré, enregistrés dans la même chambre que le corpus d'apprentissage et dans le laboratoire d'ordinateurs (la dynamique du signal était dans ce cas 30 dB env.). Tous les enrégistrement étaient effectués avec un microphone dynamique de haute qualité. Le signal était filtré passe-basse 5 kHz, échantilloné à 12 kHz et représenté par des coéfficients cepstraux calculés tous les 10 ms d'un segment du signal de 25.6 ms pondéré par une fenêtre d'Hamming. Chaque trame est représentée par un vec-teur de 28 coefficients (14 coefficients cepstraux LPC + leurs différences temporelles).

3. LES CLASSES ARTICULATOIRES

Comme on vient de dire, on a choisi telles classes articulatoires qui sont le moins sensibles aux phénômènes de coarticulation, dont leur forme physique est stable et en plus indépendante de traits personnels de la voix du locuteur, de son état émotionnel, etc.

Dans le travail précédant [4,5] on a défini dans l'espace paramètrique 4 classes articulatoires seulement: les résonants (RS), les fricatives sourdes (SS), les occlusives voisées (OC) et les occlusions sourdes (UC). Ces classes étaient définies dans l'espace de trois paramètres à mesurer, dont le niveau d'envelope et les niveaux d'énergie dans les basses et hautes fréquences. Chaque trame de 10 ms etait qualifiée comme apartenante à une de classes donées. En utilisant, en plus, de règles contextuellement dépendantes le signal de parole était décrit à l'aide de 9 classes articulatoires, en somme.

Cette fois le choix de classes articulatoires est fondé sur un système plus cohérant. Sur la Fig.1 on présente le schéma de classification de sons polonais d'après le mode, lieu et hauteur d'articulation. Le trait voisé\non-voisé n'est pas distingué et il est inclu dans le mode d'articulation. Ce trait est d'ailleur bien sensible à la coarticulation, mais les règles de voisement ou dévoisement sont assez simple. Le groupe le plus nombreux est la classe de résonants qui outre les voyelles comporte les semi-voyelles, la consonne /l/ et les nasales, c.-à-d., les sons prononcés avec une configuration du conduit vocal assurant le passage libre du soufle d'aire sur toute la longueur du conduit.

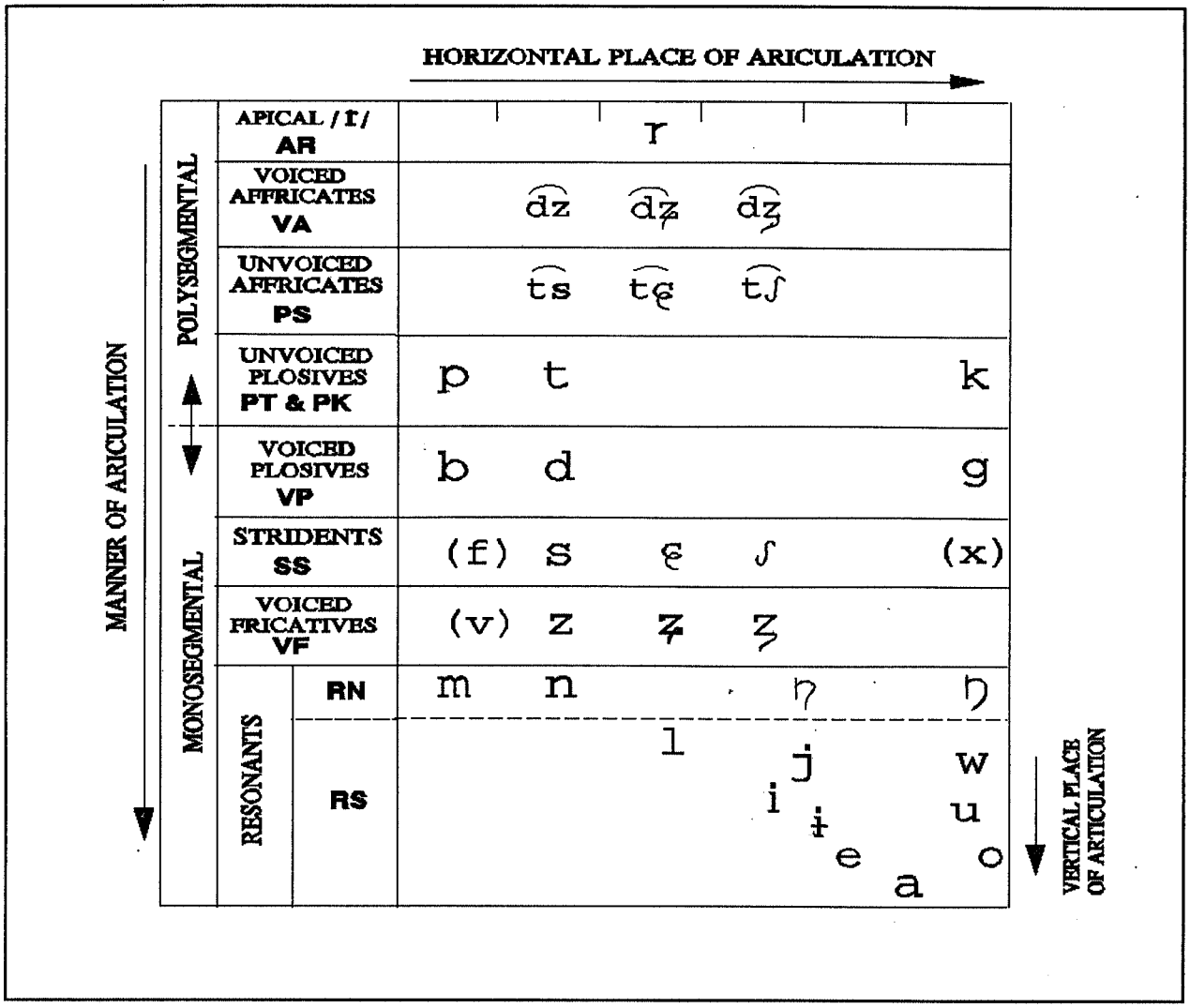


Fig. 1 Schéma de classification de sons polonais

Dans le groupe de sons qualifiés comme résonants on n'a pas inclus les voyelles nasales (en polonais il y en deux - /ã,ẽ/), car dans la plupart de cas elles sont réalisées sous forme d'une suite de deux segments distincts: un vocalique correspondant à la voyelle ouverte suivi d'un segment nasale proche de /n/.

Pour le moment, le lieu et l'hauteur d'articulation n'étaient pas utilisés et d'autres traits d'un caractère un peu plus acoustique que articulatoire étaient introduits. Néaumoins, l'interprétation articulatoire de ces traits ne pose pas de problème. Parmi les classes définies d'une façon évidente on a distingué pour les fricatives deux classes de consonnes qui malgré leur appartenance au même groupe articulatoire ont des caractéristiques acoustiques fortement différentes. Ainsi, le lieu d'articulation était remplacé par un autre trait, appelé la force d'articulation. A la base de ce trait on a distingué deux paires de consonnes fricatives non-voisées: faibles (WS) et fortes - (SS), et de même pour leur correspondantes voisées - (WF) et (VF) (sur le schéma les consonnes faibles sont prises en parenthéses). En plus, on a divisé les occlusives sourdes en deux classes nommées (PT) et (PK), dont la deuxième comporte les consonnes /p/ et /k/ qui se distinguent de la consonne /t/ par une basse explosion après l'occlusion. Comme on vient de dire, ce trait est plutôt acoustique que articulatoire, mais pour definir le lieu d'articulation de la consonne plosive il faut analyser surtout la partie transitoire du segment vocalique voisin, ce qui mènerai pra-

tiquement à l'augmentation le nombre de modèles. Et cela on a voulu éviter au stade actuel d'étude.

4. APPLICATION DE MODELES DE MARKOV CACHES

Dans l'approche par modèles de Markov cachés la parole est traitée comme un phénomène acoustique correspondant à une chaîne de phonèmes sous-jacente. Elle est donc composée d'une représentation d'une chaîne non-observable et d'un processus aléatoire qui peut être directement mesuré et qui représente la validité de la structure acoustique du processus correspondant à la chaîne de phonèmes modelisés. Dans notre cas, chaque classe articulatoire est modélisé par un modèle fonctionnant de gauche à droite, comportant 5 états avec deux particularisés: état initial et final. La sortie du premier et le passage au dernier sont des transitions vides. Pour chaque état une matrice de covariances diagonale est calculée.

Le système d'étiquetage basé sur les modéles HMM comporte deux étapes. Le premier, dite d'apprentissage, a lieu une fois à condition que le corpus utilisé pour ce but soit suffisament représentatif pour le langage utilisé, indépendament de la voix du locuteur. Le second, au cours duquel les étiquettes sont alignées sur le signal analysé, a lieu chaque fois pour décrire le signal à l'aide d'une séquence des symboles de classes articulatoires choisies.

5. L'APPRENTISSAGE ET VERIFICATION DU SYSTEME

La séquence de référence de symboles des classes articulatoires était obtenue par la conversion automatique du code orthographique en symboles phonétiques et leur classification suivante d'après le schéma présenté. La conversion était réalisée selon les règles publiées dans [7]. Outre les 12 classes articulatoires mentionées, dans la description du signal de parole deux classes de silence étaient introduites, une qui limite le signal de parole, l'autre qui se trouve à l'intérieure de la phrase analysée. Les dimensions des modèles de silence étaient differentes. Pour le modèle de la silence à l'extérieure de la phrase le nombre d'états était 7, pour la silence à l'intérieure - 5.

L'apprentissage du système a commencé sur un vocabulaire restreint à une vingtaine de mots de la liste mentionée, prononcés en plusieurs pseudo-phrases d'une façon continue par trois. Après les avoir étiqueter manuellement on a calculé les modèls de classes articulatoires et le reste du vocabulaire était étiqueté automatiquement. Ensuite, on a corrigé tous les erreurs et on a recalculé les modèles pour tout le corpus de 100 mots.

La vérification par processus de reconnaissance sur le corpus d'apprentissage a montré que l'exactitude (substitions et délétions inclus) était d'ordre de 90% et avec les insertions 85%. Il est à noté que le choix de classes articulatoires ne peut pas être tout à fait arbitraire. Par exemple, les consonnes nasales étaient au début dans la même classe que les voyelles, c.-à-d., de résonants. Le taux d'erreur pour le corpus d'apprentissage était dans ce cas de 10% plus grand qu'après la séparation de consonnes nasales en un groupe distinct.

La vérification de l'efficacité du système de description articulatoire basé sur les modèles de Markov cachés était réalisée par l'expériment d'alignement de 20 phrases phonétiquement équilibré, prononcées par plusieurs locuteurs. La durée de prononciation d'une phrase d'une longueur moyenne 6.5 mots était environ 3 s. Les phrases étaient enrégistrées dans de conditions différentes de celles qu'on a eu pour le corpus d'apprentissage, ce qui a permis d'observer l'influence du niveau de bruit sur le taux d'erreur d'alignement.

Les modèles obtenus lors d'apprentissage étaient utilisés pour l'étiquetage d'onde de parole par l'algorithme de Viterbi et l'alignement était effectué par la programmation dynamique. Pour le corpus de test enrégistré dans le même condition que le corpus d'apprentissage le taux d'erreur (substitutions et délétions) était de 4% env. plus grand (pour le même locuteur). Mais pour le corpus de test enregistré dans de conditions moins favorables avec un bruit ambiant plus important, c'est surtout le nombre d'insertions qui a augmenté d'une façon considerable (de 15%). Les insertions ont eu lieu premièrement dans la zone de silence précédante et subséquente à la phrase et deux catégories de sons y étaient surtout erronément inserrés: les fricatives sourdes faibles (WS) et les occlusives sourdes (PK). En plus, on observe relativement souvent une substitution des consonnes (PK) par les fricatives (WS). Une solution partielle à ce problème consiste

à récalculer les modèles pour ces classes de sons enregistrés dans de conditions nouvelles, ainsi que pour les silences. Avec les modèles actualisés le nombre d'insertions a baissé d'un tiers pour la classe (PK) et 40% pour la classe (WS). Ansi, le taux d'erreurs (substitutions et délétions) a baissé à 11% (avec les insertions à 22%). Les résultats cités étaient obtenus pour un locuteur. Les expériences avec un groupe de plusieurs locuteurs sont en cours et les résultats finaux seront présentés au cours du seminaire.

6. CONCLUSIONS

L'approche articulatoire en analyse et reconnaissance de parole semble d'être de plus en plus populaire [8,9]. Elle permet d'un côté d'éviter un surdéveloppement du nombre de modèles afin d'obtenir un taux d'erreurs au desssous d'un niveau acceptable, d'autre part elle est très simple et facilite l'analyse ultérieure. Elle a tous les atouts d'une description en macro-classes qui est robuste et utilise relativement peu de symboles. Un autre avantage est un nombre restreint de règles de coarticulation simples et faciles à interpréter dans le domaine acoustique. De même, les erreurs du décodage peuvent être sans équivoque liées à la forme du signal acoustique, ce qui facilite d'établir leur source et d'introduire des corrections lors d'apprentissage. Le choix de classes pour la description primaire du signal de parole doit être fait attentivement.

Les semi-voyelles posent un problème à part. Malgré que leur présence dans le signal est facile à détecter, l'évaluation de leurs limites n'est pas si évidente. C'est la raison pour lequel on a laissé les sons d'un caractére fortement transitoire (outre la consonne /r/) pour l'étude ultérieure.

REMERCIEMENTS

Le travail référé était partielement réalisé avec le support de l'Institut de Linguistique et Phonétique de l'Université de Leeds. Les modèles de Markov étaient implémentés à l'aide de HTK Toolkit reçu de M. Steve Young de CUED Université de Cambridge et adapté au PC 286 par M. A. Wrzoskowicz de L'Institut des Recherches Fondamentales de Technologie.

REFERENCES

[1] LEE K.-F. , Automatic Speech Recognition. The Development of the SPHINX System, Kluwer Ac. Pub., Boston, 1989.

[2] LJOLJE A., LEVINSON S.E., Development of an Acoustic-Phonetic Hidden Markov Model for Continuous Speech Recognition, IEEE Trans. on Acoustics, Speech Signal Processing, ASSP-39, 1, 29-39.

[3] GUBRYNOWICZ R., MARASEK K., WIĘŹLAK W., Reconnaissance de mots isolés par la méthode de traits phonétiques, 15èmes Journées d'Etudes sur la Parole, Aix-en-Provence, 1986, 235-238.

[4] GUBRYNOWICZ R., An approach to articulatory representation of speech signal on the basis of its approximate parametric analysis, Proc. of the "Speech Research'89" Conf., Budapest, 1989, 273-276.

[5] GUBRYNOWICZ R., On articulatory description rules in speech signal analysis and recognition, Proc. of the XIIth Int. Congress of Phonetic Sciences, v.3, Aix-en-Provence, 1991, 386-389.

[6] YOUNG S.J., HTK: Hidden Markov Model Toolkit V1.2, Reference Manual, Cambridge University Eng. Dept., December 1990.

[7] STEFFEN-BATOGOWA M., L'automatisation de la conversion phonématique des textes polonais (en polonais), PWN, Warszawa, 1975.

[8] DALSGAARD P., ANDERSEN O., BARRY W., The cross-language validity of acoustic-phonetic features in label-alignment, Proc. of the XIIth Int. Congress of Phonetic Sciences, v.5, Aix-en-Provence, 1991, 382-385.

[9] KABRE H., PERENNOU G., VIGOUROUX N., Automatic labelling of speech signal into phonetic events,Proc. of the XIIth Int. Congress of Phonetic Sciences, v.5, Aix-en-Provence, 1991, 450-453.

19es J.E.P. - Bruxelles, 19 au 22 mai 1992.

MODELISATION DE LA DUREE GLOBALE D'UN SON DANS UN MODELE DE MARKOV CACHE: APPLICATION A LA RECONNAISSANCE DE NOMBRES

N. SUAUDEAU, R. ANDRE-OBRECHT, B. DELYON

IRISA - URA 227 - CENTRE INRIA DE RENNES

Résumé

En reconnaissance de parole par HMM, le réseau probabiliste s'obtient hiérarchiquement. Chaque mot est décomposé sous forme d'unités linguistiques intermédiaires: une source acoustique markovienne est alors associée à chaque unité. L'inconvénient de cette modélisation est que la dimension temporelle du signal de parole n'est pas prise en compte. Afin d'y remédier, il est courant d'introduire des lois sur le temps d'occupation d'un état. L'obtention de durée raisonnable au niveau phonétique intermédiaire reste néanmoins non garantie.

Le nouveau modèle fait suite à un prétraitement original, incluant un algorithme de segmentation automatique. Une nouvelle coordonnée est adjointe au vecteur d'observations: la longueur du segment qu'il paramétrise. Il suppose qu'il existe deux niveaux hiérarchiques distincts pour la considération des observations, suivant leur nature . Le paramètre de durée est intégré au niveau du phonème, alors que les autres paramètres, de nature spectrale, restent introduits au niveau acoustique élémentaire qu'est l'état.

1 INTRODUCTION

La reconnaissance de la parole se sert couramment des modèles markoviens. En général, la représentation d'une application, sous forme d'un réseau markovien unique, résulte d'une procédure hiérarchique.

Dans un premier temps, l'ensemble des connaissances a priori conduisent à la construction des réseaux syntaxiques et lexicaux, dans lesquels tout chemin allant de l'état initial à l'état final est une phrase valide du vocabulaire traité. Puis, chaque mot est décomposé en une suite d'unités intermédiaires, possédant une nature phonétique (phonèmes, allophones, ...), qui servent de supports aux modèles acoustiques.

L'étape ultime consiste à substituer son modèle markovien à chacune des unités élémentaires. On présume alors qu'au niveau acoustique, un son correspond à un jeu de cibles acoustiques, dont chacune à une réalisation présentant des traits formantiques et énergétiques spécifiques. Ainsi, dans le HMM associé à un phonème, les probabilités de transitions entre états peuvent s'interpréter comme modélisant la coordination temporelle des différentes étapes de sa prononciation. De plus, chaque transition est liée à une des cibles acoustiques; l'observation qu'elle émet suit une loi de probabilité qui décrit la variabilité spectrale de la réalisation acoustique de cette cible.

Toutefois, une des déficiences des HMM est sa pauvreté en ce qui concerne la modélisation de la dimension temporelle de la parole. Un premier remède passe par l'utilisation de fonctions de densité de probabilité continues, modélisant le séjour dans un état. Cette alternative a été dévellopée par Levinson ([3]), et reprise par Jouvet ([2]).

De telles stratégies semblent insuffisantes. Cette étude propose d'introduire la durée à un niveau supérieur à l'état: le niveau phonétique.

Après avoir décrit l'ensemble du dispositif de reconnaissance qu'on cherche à améliorer, une série de tests sur les durées phonétiques est présentée. Puis, le nouveau modèle, ainsi que les procédures d'apprentissage et de reconnaissance associées, sont proposées avant de récapituler les résultats expérimentaux obtenus.

2 SYSTEME DE BASE

La chaîne de reconnaissance que nous proposons se compose de deux modules: un étage de prétraitement acoustique, qui extrait l'information significative du signal de parole; suivi du décodeur linguistique qui, au moyen d'un modèle markovien, interprète au mieux les observations qu'il reçoit. Sa spécificité réside dans une approche segmentale au niveau du prétraitement acoustique. Une précédente étude ([1]) a évalué l'ensemble du dispositif.

* Prétraitement acoustique

Dans les systèmes standards, le prétraitement repose sur une approche du type bloc-glissants. Une analyse fréquentielle, portant sur des trames de longueur fixe régulièrement espacées, permet de calculer les paramètres acoustiques (MFCC, LPCC, ...). En entrée du décodeur, le signal se présente comme une suite de vecteurs d'observations de dimension correspondant au nombre de coefficients acoustiques retenus par trames.

Dans l'approche segmentale, l'adjonction d'un algorithme de segmentation automatique permet d'isoler les zones homogènes du signal, grâce à une détection séquentielle des modifications spectrales du signal (elles manifestent l'existence des changements articulatoires). Ainsi, la corrélation spectrale qui existe entre des trames successives issues d'une même zone stable est exploitée; on espère aussi tirer profit d'une meilleure localisation des évènements articulatoires. L'analyse acoustique opère dès lors sur ces portions de taille variable du signal. Le jeu d'observation est constitué d'un vecteur par segment. Par rapport à l'approche bloc-glissant, chaque vecteur contient une coordonnée supplémentaire: la durée du segment qu'il paramètrise.

Les paramètres acoustiques sont, en l'occurence, les 8 premiers coefficients MFCC, auxquels on ajoute un terme d'énergie, ainsi que sa variation temporelle.

* La modélisation markovienne

Le réseau markovien est standard: il s'obtient comme déjà décrit. L'unité de base choisie est l'allophone; les effets contextuels sont ainsi pris en compte.

Il reste à définir la structure du modèle associé à chaque unité intermédiaire; soit, son nombre d'états, les transitions et les lois sur les observations, liées aux transitions. Des connaissances phonétiques a priori, sur l'articulation du son qu'il représente, sont introduites, de sorte qu'on est amené à prendre un type de modèle par grande classe phonétique. Ainsi, le modèle d'une voyelle orale (resp. nasale) comporte 4 (resp. 5) états.

La modélisation sera complètement spécifiée une fois qu'on aura associé des densités de probabilités à la chaîne de Markov. On suppose que les paramètres acoustiques sont indépendants les uns des autres. Nous contentant d'une approximation gaussienne, les densités retenues sont en fait des distributions gaussiennes, de matrices de covariance diagonales.

* La base de données

La base de données sur laquelle ce modèle a été évalué provient du CNET. Elle traite un vocabulaire relatif aux nombres de 0 à 999, dits par 70 locuteurs distincts. Au cours des expériences, elle est divisée en deux groupes: une première moitié permet d'apprendre le modèle, l'autre sert aux tests. Afin de garantir une évaluation indépendante du locuteur, aucun d'eux n'est commun aux deux subdivisions.

* Résultats

Les expériences menées montrent qu'une approche segmentale, tout en conduisant à des taux de reconnaissance similaires à ceux obtenus par analyse sur fenêtre glissante (environ 94%), diminue la quantité d'observations traitée par le décodeur linguistique.

Il convient de souligner qu'au cours de l'étude comparative, le paramètre temps était pris en compte. L'approche par bloc-glissant utilise l'algorithme de Jouvet ([2]): une loi gaussienne modèlise le temps de séjour dans chaque état. Pour ce qui est de l'approche segmentale, la durée est introduite plus immédiatement: il suffit de considérer la longueur du segment comme étant une nouvelle coordonnée du vecteur d'observations.

Dans le cas d'une approche segmentale, l'amélioration de 2%, sur le taux de reconnaissance, apportée par l'intégration de la durée, prouve la valeur informative de ce paramètre. Toutefois, il apparait que la modélisation de la durée est insuffisante, puisqu'il subsiste des reconnaissances erronées dont un exemple typique est celui de la figure 1: 81 est reconnu comme 80, d'où un allongement exagéré du phonème /ɛ̃/.

Notre objectif est donc de proposer, tout en gardant l'approche segmentale, un modèle qui considère la durée à un niveau plus global que l'état tel que le niveau phonétique. Un préambule à cette étude passe par l'acquisition de connaissances sur la durée des sons.

3 LES DUREES DES SONS

La plupart des phonéticiens admettent que la durée des sons est une source d'information interne à la parole ([4]), qui peut s'avérer essentielle à la distinction de mots acoustiquement proches. Dans cette section, on cherche à extraire des règles phonétiques significatives, ayant trait aux phénomènes temporels et susceptibles d'être incorporées au modèle de reconnaissance.

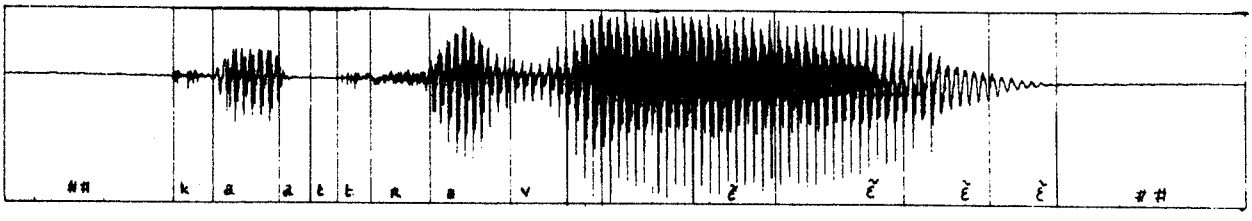


Figure 1: Un exemple de reconnaissance erronée sur le nombre "81"

Pour ce faire, l'analyse des durées des sons décrites ici se restreint au type de parole concernée par l'application. La création d'une base de données met à profit l'ensemble du dispositif de reconnaissance déjà décrit. Dans un premier temps, chaque prononciation est reconnue, ce qui conduit à la pose d'une étiquette (nom de la loi reconnue comme source) sur chaque segment. Le calcul des durées des phonèmes (seuls les résultats des reconnaissances justes sont utilisés) s'effectue alors en sommant les longueurs des segments adjacents reconnus comme émis dans un même modèle allophonique élémentaire. Ces informations, nom du phonème et sa durée, ainsi que diverses autres, qui semblent pertinentes: position du son dans le mot, phonèmes adjacents, ... sont conservées.

Les principaux points qui émergent, suite au traitement de ces données, sont les suivants.

Dans les mots monosyllabiques, chaque son semble posséder une durée inhérente, liée pour l'essentiel à sa classe phonétique. Ainsi, on retrouve que les consonnes fricatives sont généralement plus longues que les occlusives non voisées. De même, pour les voyelles, la distinction orale / nasale se constate aussi au niveau de leurs durées moyennes. Par ailleurs, des écarts plus faibles peuvent apparaitre dans le cas de sons tirés d'une même catégorie phonétique mais ayant des lieux d'articulations distincts (cf cas des voyelles hautes /i/ plus courtes que les voyelles basses /a/).

Il convient de souligner qu'ici, la notion de "son plus court" ne se réfère pas uniquement à une durée de valeur moyenne plus faible; elle traduit aussi le fait qu'à la fois la durée minimum observée (dont l'existence démontre l'impossibilité de compresser l'articulation d'un son en dessous d'une limite) et la durée maximum présentent des valeurs plus petites.

Cependant, dans les mots à plusieurs syllabes, la durée du son devient très variable: elle subit de multiples influences qui agissent comme autant de facteurs, aux effets à la fois multiplicatifs et additifs, sur sa valeur inhérente.

Au niveau lexical, le mot dont est issu le son, et surtout sa position dans ce mot, sont non négligeables. L'appartenance à la dernière syllabe conduit à allonger le son, conséquence probable de la proximité du silence. Un exemple caractéristique est celui de la dernière voyelle du mot, dont la durée minimale dépasse systématiquement 100 ms, alors que pour toutes autres positions, les voyelles prennent des durées minimales moindres.

Au niveau phonétique, des effets contextuels importants apparaissent (nota: bien que présents dans les mots monosyllabiques, l'existence d'un seul contexte possible pour la plupart des sons masquaient ces influences). Par exemple, pour l'occlusive /t/ située après une voyelle nasale, sa durée moyenne vaut 30 ms, avec un maximum possible de 100 ms quand elle précède la fricative /s/. Mais, si elle est suivie par une des voyelles /yi/,/e/,/i/, elle s'allonge très nettement: sa durée moyenne excède 100 ms, et elle ne dure jamais moins de 50 ms. Plusieurs justifications sont possibles: difficultés articulatoires de certaines combinaisons phonétiques, cibles acoustiques non atteintes, puisque d'autres sources, à un niveau de connaissances supérieur, permettent de faire une distinction, ...

On retient de cette analyse que la durée d'un son est bien corrélée à son identité, ainsi qu'à son contexte; l'intégrer au modèle de reconnaissance doit permettre de corriger certaines erreurs.

La stratégie retenue revient à admettre que la durée de chaque entité phonétique est une variable aléatoire possédant, de manière absolue, une loi de probabilité. Une restriction est apportée à l'hypothèse précédente, pour les sons dont la durée est soumise à d'importantes fluctuations, suivant son voisinnage, sa position, ... ; on veille à leur attribuer autant de lois qu'il y a de contextes influents.

4 MODELE A DEUX NIVEAUX

Le nouveau modèle exploite la possibilité offerte par l'approche segmentale: la longueur du segment est une des coordonnées du vecteur d'observation. Tout en gardant une structure de réseau markovien, il utilise deux niveaux hiérarchiques distincts pour la considération des observations, suivant la nature des renseignements qu'elles contiennent. Le niveau phonétique, qui est supérieur aux états, sert à traiter le paramètre global qu'est la durée, tandis qu'au niveau local des paramètres liés à chaque état du modèle acoustique sont introduits.

Dans le cas standard, le modèle probabiliste permet d'évaluer la vraisemblance qu'à un chemin $s_T = q_{i_1}, \ldots, q_{i_T}$ d'avoir engendré une suite d'observations $O = O_1, \ldots, O_T$, grâce à deux catégories de lois: les lois de probabilités de transitions entre états du réseau $a_{i,j}$, ainsi que les lois sur les observations, liées aux états $b_i(O_t)$.

Ici, les observations sont toujours présumées engendrées par un chemin du réseau: réseau dont la structure est inchangée par rapport au modèle classique (mode d'obtention identique). On postule donc l'existence de lois $a_{i,j}$ qui régissent les transitions entre états.

En outre, chaque état fait partie d'un modèle allophonique par construction. D'où, à tout chemin s_T, il correspond, de façon déterministe, au niveau phonétique, une suite:

$$\varphi(s_T) = \begin{pmatrix} phon \\ ptr \end{pmatrix}$$

$$avec : \quad \begin{aligned} phon &= \phi_{j_1}, \ldots, \phi_{j_{mx}} \\ ptr &= d_1, \ldots, d_{mx} \end{aligned}$$

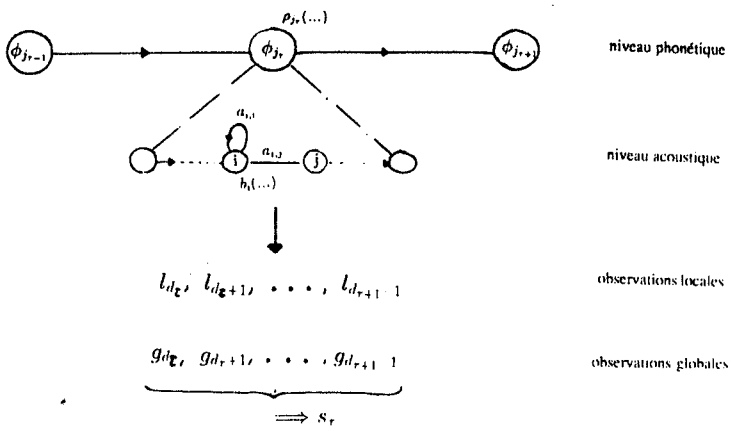


Figure 2: Vue hiérarchique du modèle à deux niveaux

Quand on parcourt le chemin s_T:

- mx est le nombre total d'allophones rencontrés
- ϕ_{j_i} est le nom du $i^{ème}$ allophone atteint au niveau phonétique
- d_i est un pointeur dans ce chemin, indiquant l'indice temporel du 1^{er} état appartenant au $i^{ème}$ allophone.

Pour avoir totalement défini le modèle, il faut choisir des lois de probabilités sur les observations. Ces dernières sont décomposées en deux suites: les observations globales $G = G_1, \ldots, G_T$ et les observations locales $L = L_1, \ldots, L_T$.

G_i représente la longueur du segment paramétrisé par le vecteur O_i et L_i se compose des coordonnées restantes du vecteur O_i, autres que sa durée.

Pour les observations locales L, on retrouve les hypothèses du modèle standard. L'observation L_t dépend seulement de son état source q_{i_t}. Elle est présumée émise suivant une loi $b_{i_t}(..)$.

En ce qui concerne les observations globales G, on suppose qu'elles sont indépendantes des observations locales. Par ailleurs, elles ne sont pas directement liées aux états, mais uniquement à la suite phonétique $\varphi(s_T)$. Plus précisement, une observation G_t est supposée dépendre du phonème courant, ainsi que des autres observations globales adjacentes émises dans ce même son. Ce qui revient à écrire:

$$P(g_1, \ldots, g_T/\varphi(s_T)) = \prod_{i=1}^{mx} P(g_{d_i}, \ldots, g_{d_{i+1}-1}/\phi_{j_i})$$

Le calcul des termes du produit ci-dessus nous oblige à faire deux hypothèses supplémentaires. La première concerne la répartition du temps de séjour dans le $\tau^{ème}$ phonème, soit $s_\tau = g_{d_\tau} + \cdots + g_{d_{\tau+1}-1}$, entre les $(d_{\tau+1} - d_\tau)$ segments acoustiques allignés sur ce phone. On admet l'équiprobabilité de toutes les combinaisons possibles $(g_{d_\tau}, \ldots, g_{d_{\tau+1}-1})$ conduisant à un même temps de séjour s_τ. De plus, le temps total passé dans le son ϕ_i est modélisé par une loi $\rho_i(...)$. D'où:

$$P(g_{d_\tau}, \ldots, g_{d_{\tau+1}-1})/\phi_{j_\tau}) = [\rho_{j_\tau}(s_\tau) \times (d_{\tau+1} - d_\tau - 1)!]/s_\tau^{(d_{\tau+1}-d_\tau-1)}$$

5 ALGORITHME DE RECONNAISSANCE

L'utilisation du modèle à deux niveaux préserve le principe de recherche d'un meilleur chemin dans le réseau, comme fondement de l'algorithme de reconnaissance. Toutefois, sa perte de la propriété markovienne, au sens strict, oblige à introduire un indice supplémentaire dans la procédure de Viterbi.

Dans le cas standard, l'algorithme se contente de conserver, à chaque itération t, pour chaque état du réseau, le chemin partiel s'y terminant qui a la plus forte probabilité d'avoir émis $O_1, \ldots, O_t$; il est le seul qui soit susceptible, parmi l'ensemble des chemins partiels arrivant à cet état, d'être le début du chemin optimal de longueur T recherché.

A l'instant t, le nouveau modèle rend impossible le calcul de la probabilité qu'à un chemin partiel d'avoir émis les observations $l_1, \ldots, l_t$ et $g_1, \ldots, g_t$. Tout au plus, sous réserve que le chemin partiel condidéré soit entré dans le dernier phonème courant depuis sej observations, on peut calculer sa vraisemblance d'être la source de $l_1, \ldots, l_t$ et $g_1, \ldots, g_{t-sej}$.

Or, la pénalisation supplémentaire apportée sur sa vraisemblance, lors de sa prolongation future provient: du coût des transitions $q_{i_t} \rightarrow q_{i_{t+1}} \rightarrow \ldots \rightarrow q_{i_T}$, de l'émission des observations locales $l_{t+1}, \ldots, l_T$ sur ce chemin, ainsi que de l'émission des observations globales $g_{t-sej+1}, \ldots, g_T$.

D'où, à un instant t, soit deux chemins partiels arrivant à un même état, mais correspondant à des instants d'entrée dans le phonème courant distinct: bien qu'em-

pruntant, par la suite, les même états, il est impossible de savoir lequel des deux conduira, après totale prolongation, à la plus forte probabilité d'émission de l'ensemble des observations. Par contre, s'ils ont déjà passé le même temps dans le phonème courant, une prolongation semblable conduit à une même pénalisation; il est donc inutile de conserver le moins bon.

C'est pourquoi, pour chaque état, on garde un chemin partiel par instant possible d'entrée dans le son correspondant. Chacun d'eux est indicé par la grandeur sej: il est, parmi l'ensemble des chemins se terminant, à l'instant t, dans l'état q_i, après un séjour de sej observations dans ce dernier son, celui qui a le plus vraisemblablement engendré les suites $l_1, \ldots, l_t$ et $g_1, \ldots, g_{t-sej}$.

Afin de déterminer les chemins à conserver, les formules de réccurence sont du même type que dans le cas classique. Toutefois, on distingue deux groupes d'états: ceux qui sont des fins de phonèmes et les autres. Puis, pour un état, suivant la valeur prise par le nouvel indice sej, on recherche la prolongation optimale, soit d'un chemin arrivant d'une fin de son (cas $sej = 1$), ou bien d'un chemin étant déjà dans ce son depuis $sej - 1$ observations (cas $sej \geq 1$). C'est au niveau des états de sortie de sons qu'on retient un seul chemin partiel: celui qui est le meilleur, une fois introduites toutes les dernières observations de durées reçues.

La phase d'apprentissage reprend les méthodes traditionnelles. Une procédure itérative, dérivée de l'algorithme de Baum-Welch, permet d'optimiser les paramètres du modèle, à partir de reconnaissances et de réestimations successives. A chaque itération, le calcul des nouveaux paramètres prend en compte uniquement les évènements (transitions, observations dans un état, un phonème, ...) relatifs à l'alignement d'une observation sur le chemin optimal reconnu comme sa source.

6 PREMIERS RESULTATS

L'évaluation des performances de cette nouvelle modélisation utilise l'application traitant de la reconnaissance des nombres. Le dispositif de reconnaissance de la section 2 est donc repris. Aucune modification n'est apportée au prétraitement acoustique. Par ailleurs, le modèle conserve une description à base d'allophones, dans laquelle nous distinguons, pour un son donné, les contextes qui, au cours des tests sur les durées, sont apparus comme influençant significativement sa longueur. Puis, l'allophone est représenté par un modèle acoustique, possédant une structure compatible avec les résultats de la segmentation, de sorte qu'elle est liée à sa classe phonétique.

Dans une première phase de la mise en oeuvre, la compilation du réseau acoustique final de l'application, ainsi que l'apprentissage des paramètres du modèle markovien classique équivalent, où les lois de durée dépendent de l'état (cf section 2), utilisent les logiciels développés par le CNET. Les résultats obtenus après apprentissage sont repris afin d'initialiser les paramètres du nouveau modèle, excepté pour les lois sur les durées des sons. L'initialisation de ces dernières n'est pas aussi directe: les grandeurs statistiques qu'elle nécessite (moyennes, variances, ..) sont obtenues grâce aux résultats de reconnaissance du modèle équivalent (idem section 3). Pour finir, les paramètres du nouveau modèle sont optimisés, au moyen de la procédure d'apprentissage précédemment décrite, qui nécessite un très faible nombre d'itération (seulement 2 à 3).

Trois catégories de distributions de probabilités, modélisant la durée d'un son, ont été comparées:

* Des lois gaussiennes, qui sont définies par les paramètres de moyenne et de variance. Leur intérêt est que les estimateurs empiriques de ces grandeurs sont aussi ceux qui maximisent la vraisemblance.

* Des distributions Gamma, utilisées par Levinson, qui sont paramétrées par ν_ϕ et η_ϕ:

$$\begin{cases} \rho_\phi(x) = 0 & x \leq \gamma_\phi \\ \rho_\phi(x) = (x-\gamma_\phi)^{(\nu_\phi - 1)} & \times \frac{\eta_\phi^{\nu_\phi}}{\Gamma(\nu_\phi)} \times e^{[-(x-\gamma_\phi)\eta_\phi]} \end{cases}$$

La moyenne et la variance de telles lois valent respectivement $\nu_\phi/\eta_\phi + \gamma_\phi$ et ν_ϕ/η_ϕ^2. Deux cas, suivant les valeurs prises par γ_ϕ, vont être étudiés: soit le seuil est nul et $\gamma_\phi = 0$, ou bien $\gamma_\phi = 0.9 \times dmin_\phi$, avec $dmin_\phi$ correspondant à la durée minimale observée pour ce son dans l'ensemble d'apprentissage.

Il reste alors à estimer deux paramètres : η_ϕ et ν_ϕ, pour avoir complètement spécifié la loi $\rho_\phi(..)$. Les formules de réestimations sont similaires quel que soit le type de seuils. La difficulté essentielle provient de la non linéarité d'une des formules de réestimation, selon le maximum de vraisemblance, ce qui oblige à utiliser un algorithme de type Newton pour la résoudre.

* Une dernière alternative consiste à représenter les durées phonétiques au moyen d'inverses gaussiennes (idem loi de Wald), paramétrées par μ_ϕ et λ_ϕ.

Par définition, une variable aléatoire x suit une distribution inverse gaussienne si elle présente une fonction de densité de probabilité de la forme ([5]):

$$\begin{cases} \rho_\phi(x) = 0 & x \leq 0 \\ \rho_\phi(x) = \sqrt{\frac{\lambda_\phi}{2\pi x^3}} & \times e^{\frac{-\lambda_\phi (x-\mu_\phi)^2}{2\mu_\phi^2 x}} \end{cases}$$

La valeur moyenne de x vaut alors μ_ϕ et sa variance $\frac{\mu_\phi^3}{\lambda_\phi}$. Quelques cas particuliers de cette fonction de probabilité sont montrés à la figure 3.

L'usage de cette loi est simple. Les formules de réestimation de ses deux paramètres λ_ϕ et μ_ϕ, suivant un critère de maximum de vraisemblance restent linéaires; elles utilisent la moyenne des échantillons, ainsi que celle des inverses des échantillons.

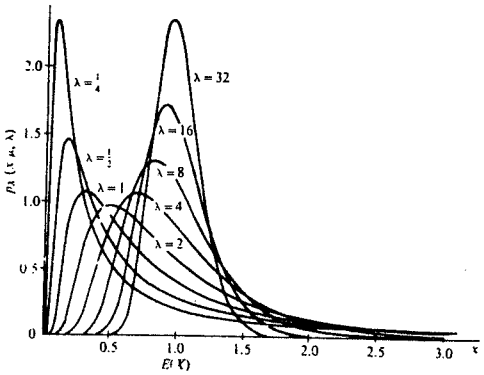


Figure 3: Fonctions de densité pour une loi de Wald

$$E(x) = \mu = 1$$

Le tableau 1 résume les performances obtenues pour le nouveau modèle, en fonction du type de lois de probabilités modélisant le séjour dans un son.

Il apparait, au niveau de l'ensemble test, qu'introduire la durée à un niveau plus global que l'état apporte un gain sur le taux de reconnaissance, allant de 0.6 % à 1.2 %, suivant le type de loi considérée.

L'importance de l'amélioration dépend du degré d'intégration de la réalité acoustique de la parole, lors du choix des densités de probabilités. Le taux de reconnaissance est naturellement plus élevé avec des lois Gamma, ou des inverses gaussiennes , à supports positifs, qu'avec des lois gaussiennes. De même, pour les lois Gamma, l'introduction d'une durée minimale propre à chaque son s'avère bénéfique. Les meilleurs performances sont obtenues avec les lois inverses gaussiennes, très certainement parce qu'avec seulement deux paramètres, elles conduisent à une assez grande diversité de répertitions possibles (cf figure 3), capable de représenter la forte variabilité du paramètre durée.

Densité de probabilité sur la durée	Ensemble d' apprentissage	Ensemble test
référence*	96.7 %	93.1 %
gaussienne	96.0 %	93.7 %
gamma à seuils nuls	96.0 %	93.8 %
gamma à seuils= dmin	95.9 %	94.1 %
inverse gaussienne	96.0 %	94.3 %

*: Durée prise au niveau du segment

Tableau 1: Taux de reconnaissance pour un modèle par allophones, suivant la loi de durée retenue

7 CONCLUSION

La modélisation de la durée globale d'un son, dans un modèle markovien, semble intéressante: elle conduit à une amélioration des performances dans le cas de son application à la reconnaissance des nombres, indépendament du locuteur. Ces résultats sont d'autant plus encourageant qu'ils sont obtenus sans tenir compte de la variabilité de la vitesse d'élocution. Or, celle-ci joue très certainement un rôle important: une étude est en cours afin de l'introduire dans le modèle à deux niveaux. Mais surtout, le nouveau modèle offre un support théorique et rigoureux pour l'étude de nombreux paramètres qui devraient être introduits à des niveaux intermédiaires, autres que le niveau acoustique de type centiseconde, tout en gardant la puissance des modèles markoviens.

REFERENCES

[1] R. ANDRÉ-OBRECHT, Reconnaissance automatique de parole à partir de segments acoustiques et de modèles de Markov cachés. *JEP de Montréal*,mai 1990.

[2] D. JOUVET, Reconnaissance de mots connectés indépendamment du locuteur par des méthodes statistiques. *thèse de Docteur, ENST*,juin 1988.

[3] S. E. LEVINSON, Continuously variable duration hidden Markov models for automatic speech recognition. *Computer Speech and Language*,1986.

[4] D. O'SHAUGHNESSY, A study of french vowel and consonant duration. *Journal of Phonetics 9*,1981.

[5] JOHNSON, KOTZ, Continuous Univariate Distributions 1. *Wiley-interscience publication*,1970.

19es J.E.P. - Bruxelles, 19 au 22 mai 1992.

RECONNAISSANCE ANALYTIQUE DE MOTS ISOLES D'UN GRAND LEXIQUE

MELONI H., BECHET F., GILLES P.

LABORATOIRE D'INFORMATIQUE UNIVERSITE D'AVIGNON

Résumé

Le système proposé permet la reconnaissance analytique de mots équiprobables d'un lexique extensible de plus de 22.000 entrées. L'adaptation au locuteur – effectuée à partir de quelques énoncés spécifiques – sélectionne des références spectrales au moyen desquelles les règles et les procédures d'identification sont ajustées. Une phase de décodage acoustico-phonétique ascendant produit un treillis d'unités phonétiques valuées qui constitue les données du processus de tri ascendant d'un ensemble de mots vraisemblables. le classement des hypothèses est fondé sur une mesure de la densité de recouvrement du signal par les références des unités phonétiques décrivant les mots probables. La cohorte d'une vingtaine de mots candidats est ensuite reclassée au moyen d'une phase descendante de décodage acoustico-phonétique qui évalue contextuellement les traces acoustiques de phénomènes articulatoires précis (propriétés résistantes de certains indices de traits phonétiques dans des situations déterminées). Les résultats sont très encourageants et permettent d'envisager à court terme un système opérationnel fondé sur cette méthodologie.

1- INTRODUCTION

Les difficultés rencontrées dans la réalisation d'un système de reconnaissance des mots d'un grand vocabulaire sont liées aux conditions d'utilisation et de fonctionnement de ce mode de communication homme-machine (Calliope, 1989). Les contraintes que nous envisageons doivent permettre :

- d'effectuer une adaptation très rapide au locuteur,
- d'enrichir de manière simple et immédiate un vocabulaire comportant plusieurs dizaines de milliers de mots,
- de réaliser la reconnaissance avec équiprobabilité de tous les mots,
- de fonctionner dans le cadre d'une énonciation naturelle de mots isolés dans un environnement peu bruité,
- d'identifier et de classer une dizaine d'hypothèses.

Afin que le système soit très rapidement opérationnel pour un locuteur quelconque, il est nécessaire d'effectuer une adaptation automatique à partir d'un ensemble limité d'énoncés définis à cet effet. La technique employée doit permettre d'extraire quelques caractéristiques essentielles du locuteur qui seront utilisées pour ajuster les règles et les processus d'identification des unités phonétiques. L'optimisation du codage et de l'emploi de ces connaissances constitue le problème majeur à résoudre.

L'extension rapide d'un important lexique (ajout immédiat d'un mot nouveau) impose une représentation synthétique sous la forme d'une séquence normalisée d'unités phonétiques. Le vocabulaire utilisé comporte 22.000 entrées (BDLEX : Pérennou, 1986), et de nombreux mots sont donc phonétiquement très proches (paires minimales). Pour tester les limites de notre méthodologie, nous avons placé le système dans la situation difficile – peu probable en pratique – où tous les mots peuvent être énoncés à chaque opération de reconnaissance. Les obstacles décisifs concernent l'accès rapide à un sous ensemble raisonnablement limité du vocabulaire et l'évaluation précise des solutions les plus probables.

Les techniques fondées sur une approche globale – relative au mot ou à des unités plus petites (Gauvain, 1986) – donnent des résultats très satisfaisants dans de nombreux contextes (systèmes monolocuteurs). Les méthodes efficaces d'adaptation au locuteur (Choukri, 1986 ; Shikano, 1986 ; Bonneau, 1987) ne sont cependant pas envisageables pour un système utilisant un large vocabulaire évolutif. Les systèmes fondés sur une approche statistique (modèles de Markov cachés) produisent des résultats remarquables lorsqu'un modèle de langage est associé au décodage phonétique pour des mots d'un énoncé séparés par des pauses (Shichman, 1986 ; Baker, 1989). Il est toutefois assez difficile, dans le cadre d'un apprentissage automatique des connaissances acoustico-phonétiques, de prendre en compte des variations pertinentes très fines du signal utilisables pour distinguer des paires minimales (traces acoustiques de phénomènes articulatoires).

Les travaux que nous avons précédemment menés pour le décodage acoustico-phonétique (Méloni, 1983, 1986, 1991a ; Bulot, 1987) et la reconnaissance multilocuteurs de vocabulaires restreints (Méloni, 1991b), nous ont conduits à utiliser des techniques basées sur une identification analytique des unités phonétiques en codant de manière explicite l'ensemble des connaissances acoustiques, phonétiques, phonologiques et lexicales. Le système présenté fonctionne en plusieurs étapes suivant le schéma de la figure 1. Les résultats obtenus pour chacune des phases de l'identification des unités (sons et mots) peuvent être utilisés de différentes façons en vue de la réalisation de systèmes de reconnaissance (mots isolés, parole continue, word spotting...).

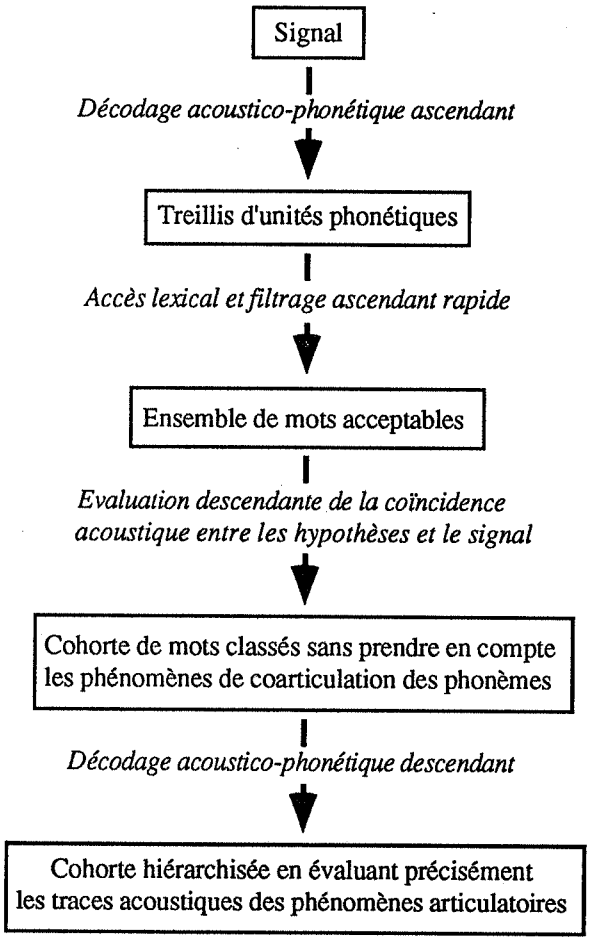


Figure 1 : Schéma général de fonctionnement du système de reconnaissance.

2- PARAMETRISATION DU SIGNAL ET ADAPTATION AU LOCUTEUR

Le signal de parole est numérisé sur 16 bits à une fréquence de 12,8 kHz puis préaccentué et caractérisé chaque 10 ms par son énergie globale, la densité des passages par zéro (signal et sa dérivée) et les énergies spectrales dans 24 canaux répartis suivant une échelle de Mel. Un ensemble d'outils permet de définir et de calculer dynamiquement de nombreux paramètres auxiliaires obtenus par combinaisons des attributs initiaux (Méloni, 1986 ; Bulot, 1987).

Les règles et les processus de décodage acoustico-phonétique – dans la phase ascendante aussi bien que dans l'étape descendante – utilisent des références paramétriques qui caractérisent certaines portions significatives des sons d'un locuteur. Ces modèles permettent d'une part, de mesurer de diverses manières la ressemblance globale du signal avec les unités phonétiques et, d'autre part, de fournir des informations acoustiques et phonétiques (position des formants, distribution spectrale de l'énergie, etc.) utilisables dans les règles décrivant les connaissances.

Les références sont sélectionnées de manière automatique dans des énoncés choisis pour présenter les phonèmes dans des situations conformes à celles qui seront proposées lors de la reconnaissance (mots isolés, mots enchaînés, parole continue). Les contextes phonétiques particulièrement déformants sont évités afin de disposer de représentants peu altérés. Lorsque l'évaluation de certains paramètres complexes (formants par exemple) ne peut être effectuée de manière déterministe, le système sollicite l'intervention de l'utilisateur pour qu'il propose une nouvelle référence moins ambiguë.

3- DECODAGE ACOUSTICO PHONETIQUE ASCENDANT

Une première phase de décodage acoustico-phonétique ascendant (Méloni, 1991a), fournit un treillis d'unités phonétiques valuées qui contient les sons essentiels (voyelles, consonnes intervocaliques fermées) permettant d'accéder à un sous ensemble raisonnablement limité du lexique. Cette étape de la reconnaissance – fondée sur des comparaisons entre les paramètres spectraux issus du signal et des références associées aux phonèmes – ne prend pas en compte les phénomènes de coarticulation et propose toutes les hypothèses phonétiques vraisemblables.

La localisation des zones coïncidant avec les noyaux consonantiques et vocaliques est effectuée – au moyen d'outils pour la reconnaissance de formes – à partir de paramètres lissés caractérisant divers types d'énergies du signal (énergie totale, énergie associée à certaines zones spectrales, etc.). Les intervalles temporels étiquetés (vocalique ou consonantique) ainsi déterminés contiennent toutes les voyelles d'un mot et, dans la plupart des cas, la consonne la plus fermée des séquences de consonnes en position intervocalique. Les cas de consonnes intervocaliques non repérées correspondent à des situations où la séquence consonantique est constituée d'un seul phonème peu fermé au contact de voyelles fermées ou de /ə/.

Les zones vocaliques sont ensuite segmentées en portions où certaines fonctions d'instabilité spectrale du signal passent par un minimum. Nous retenons, comme hypothèses phonétiques probables sur ces intervalles, toutes les voyelles dont la "distance" au signal est également minimale dans ces limites temporelles. Les fonc-

tions qui définissent les distances utilisées prennent en compte la structure particulière du phonème concerné ainsi que diverses informations contextuelles contenues dans les trames adjacentes du signal (variation de l'énergie, nouvelles fonctions d'instabilité spectrale liées au phonème, etc.).

Le score associé à chaque hypothèse est calculé au moyen d'une distance spectrale au 'son de référence sur la zone de plus grande stabilité du signal (minimum de stabilité et quelques trames adjacentes). Tous les sons qui satisfont aux conditions définies ci-dessus sont retenus, indépendamment de la valeur du score (Méloni, 1991a).

Pour les consonnes, la localisation et l'évaluation du score de vraisemblance des hypothèses sont effectuées en tenant compte du mode articulatoire du phonème (occlusif, constrictif, nasal, voisé, etc.). Cette opération n'est pas fondamentale dans la mesure où la recherche des consonnes est effectuée par la suite de manière descendante. A l'issue de cette phase, le treillis contient tous les sons vocaliques effectivement énoncées et la plupart des sons associés aux phonèmes les plus fermés des groupes consonantiques (figure 2).

4- SELECTION DES COHORTES DE MOTS

Le système de reconnaissance d'une cohorte hiérarchisée de quelques dizaines de mots utilise les données fournies par le DAP ascendant. Il effectue un tri des informations acoustico-phonétiques disponibles, opère une sélection d'un ensemble de mots compatibles avec les unités retenues et enfin calcule – pour toutes les hypothèses lexicales – le score global de ressemblance avec le signal.

4.1- Sélection des informations utiles du treillis

Le treillis est segmenté en zones de sons consonantiques et de sons vocaliques ; les zones de sons consonantiques représentent une ou plusieurs unités consécutives, les zones de sons vocaliques par contre peuvent recouvrir une série de voyelles séparées par des consonnes non obligatoirement trouvées par le DAP. Le treillis ainsi réduit définit le nombre minimal de sons vocaliques du mot prononcé.

4.2- Accès à un sous ensemble du lexique

L'accès aux mots probables du vocabulaire est effectué au moyen d'une série de filtres limitant progressivement le nombre d'items possibles du lexique général. Le dictionnaire a été partitionné en fonction de critères qui mettent en avant la première voyelle du mot, le nombre de zones voyelles obligatoires, le nombre de zones voyelles facultatives. Le premier filtrage consiste à sélectionner les différents sous-lexiques en fonction des informations obtenues dans les deux étapes précédentes. Cette opération permet de réduire de manière simple et rapide le nombre de mots possibles. En moyenne, le dictionnaire de 22000 mots est réduit à quelques milliers d'items.

Le deuxième filtre utilise l'ensemble du treillis réduit. Les mots retenus sont rangés dans un arbre en partie commune selon leur décomposition phonétique réduite aux sons obligatoirement trouvés par le DAP. Le système parcourt cet arbre parallèlement aux familles du treillis réduit en ne tenant pas compte de la place des sons à l'intérieur des familles. Le but de ce filtre est de réduire rapidement et de façon grossière le lexique en ne gardant que les mots compatibles avec le treillis conformément aux règles du DAP. Ce deuxième filtre limite généralement la cohorte à un millier de mots.

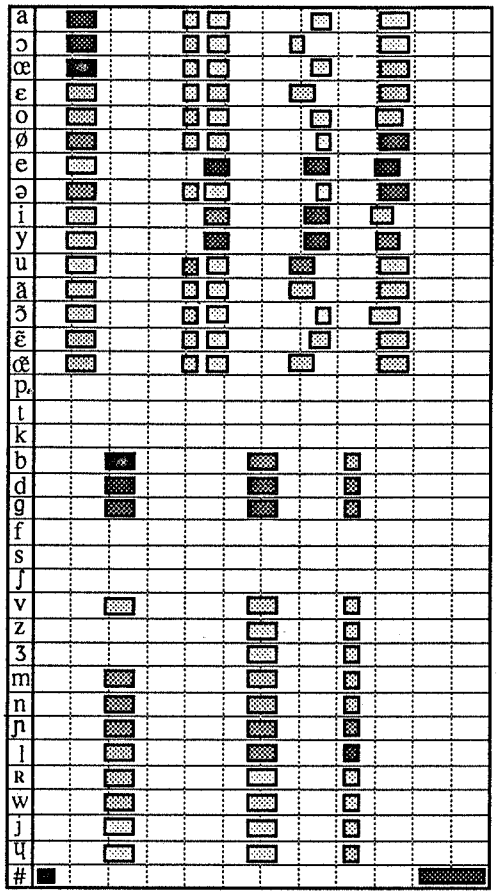


Figure 2 : Treillis phonétique obtenu après la phase ascendante de décodage acoustico-phonétique pour le mot "obnubiler". Le score de vraisemblance des unités croît avec le caractère sombre de la représentation.

4.3- Calcul du score des hypothèses lexicales

Une fois la phase de filtrage terminée, les phonèmes des mots retenus sont repositionnés sur le signal en fonction des distances spectrales aux références du locuteur. Les sons vocaliques doivent tous être présents dans le treillis ; ils sont donc placés en premier. Les sons consonantiques obligatoires sont ensuite mis en correspondance avec les zones associées du signal. Si dans cette phase un son n'est pas identifié dans le treillis, le mot est immédiatement abandonné. Les consonnes facultatives manquantes sont alors recherchées ; elles sont localisées sur la trame correspondant au minimum

de la distance à la référence du son dans la zone où il est normalement attendu.

L'évaluation de la ressemblance globale entre le signal et la projection des mots permet d'affecter une note à chaque hypothèse lexicale. La comparaison est effectuée au moyen d'une "distance" entre les références et le signal, de manière à couvrir l'essentiel de l'énoncé à l'exception des zones très instables de quelques centisecondes entre les sons qui ne sont pas prises en compte pour le calcul du score. La cohorte résultat est constituée des cinquante mots les mieux classés.

5- VÉRIFICATION DESCENDANTE FINE

La dernière étape de la reconnaissance consiste à reclasser les hypothèses au moyen d'une analyse descendante fine qui tient compte des phénomènes de coarticulation négligés dans la phase ascendante du décodage acoustico-phonétique. Il s'agit d'examiner chacune des hypothèses proposées par le système en calculant un score fondé sur la détection des phénomènes articulatoires susceptibles de distinguer les candidats.

Pour chaque phonème, nous avons décomposé sa description sous la forme d'une matrice de traits distinctifs conforme aux caractérisations habituelles des phonéticiens. Pour tous les phonèmes d'un mot, nous cherchons à identifier – dans la zone correspondante du signal – les traces acoustiques (propriétés) de certains indices caractérisant les traits de mode et de lieu articulatoires. Lorsqu'un trait est détecté positivement, il contribue à augmenter la vraisemblance de l'hypothèse lexicale traitée.

Les règles et procédures qui décrivent et identifient les propriétés dépendent du contexte phonémique utilisable dans cette étape descendante. Aucun seuil n'est défini a priori pour valider la présence d'un phénomène acoustico-phonétique, les décisions sont effectuées au moyen de valeurs qui sont dynamiquement calculées à partir des références du locuteur. Ces paramètres définissent la situation médiane – au delà de laquelle la décision pourrait être erronée – à partir des caractéristiques du son à identifier et de celles des concurrents les plus proches pour le trait concerné.

Cette phase est actuellement en cours de développement, mais les traitements qui sont aujourd'hui disponibles (voyelles, occlusives) ont montré l'efficacité de cette technique pour discriminer des paires minimales (Méloni, 1991c). La simple réorganisation de la cohorte au moyen des traits évalués sur les seules voyelles a permis d'améliorer significativement la reconnaissance. Toutefois, les progrès les plus importants sont attendus dans le traitement des consonnes qui sont beaucoup plus sensibles aux phénomènes de coarticulation et que l'on peut difficilement caractériser au moyen de références.

6- RÉSULTATS

Pour chaque phase du processus de reconnaissance, les résultats obtenus sont utilisés dans le cadre de divers types de systèmes de segmentation et d'étiquetage, d'identification du locuteur (Bonastre, 1991), de reconnaissance de mots isolés, de reconnaissance de parole continue (Spriet, 1990), de word spotting, etc. On peut donc évaluer les performances des différentes étapes pour chacune de ces tâches. Les résultats que nous donnons ci-dessous ont été obtenus sur quelques milliers de tests correspondants à l'énonciation par 4 locuteurs de mots isolés tirés au hasard parmi les 22.000 entrées du lexique BDLEX.

Le décodage acoustico-phonétique ascendant doit permettre de localiser et de mesurer un score de vraisemblance pour tous les sons d'un énoncé placés dans les situations suivantes :

- toutes les voyelles dans tous les contextes,
- la consonne la plus fermée d'un groupe consonantique situé entre 2 voyelles, à l'exception des groupes ne contenant pas de phonème sourd ou très fermé et placés au contact d'une voyelle fermée.

Pour cette opération, le système localise et identifie les sons obligatoires – avec un score convenable – dans 99,5% des cas. Moins de 2% des mots ne peuvent être sélectionnés dans le lexique à cause de l'absence d'une hypothèse phonétique utilisable.

La sélection ascendante des cohortes propose quelques dizaines d'hypothèses. Les résultats dépendent du nombre d'items retenus : 72% de reconnaissance du mot si on se limite au premier candidat et jusqu'à 97% d'identification correcte parmi les 50 meilleurs candidats (figure 3).

La réorganisation de la cohorte des mots candidats au moyen du décodage acoustico-phonétique descendant est effectuée actuellement en pondérant exclusivement les valeurs de la distance aux voyelles. Pour chaque trait dont une trace acoustique d'un des indices est détectée dans le signal, le score de ressemblance au phonème est ajusté. Une nouvelle note est globalement assignée à chaque mot. L'amélioration est peu importante dans ce cas, mais il semble cependant que les expériences effectuées sur les consonnes occlusives sourdes (Méloni, 1991c) nous permettent d'attendre un progrès significatif lorsque tous les phonèmes seront traités – notamment les consonnes très sensibles aux effets de la coarticulation.

7- CONCLUSION

Le système que nous proposons permet de réaliser un décodage acoustico-phonétique ascendant de la parole avec une adaptation très rapide au locuteur, de sélectionner une cohorte restreinte de mots d'un grand vocabulaire dans laquelle apparaît l'unité lexicale énoncée, de réordonner la cohorte au moyen d'informations contextuelles très précises concernant les traces acoustiques de phénomènes articulatoires. les résultats obtenus sont utilisables à chaque étape et ont conduit à l'élaboration de plusieurs systèmes de reconnaissance.

Une amélioration importante des résultats peut être attendue dans la dernière phase du processus lorsque le décodage acoustico-phonétique descendant concernera l'ensemble des consonnes. Il sera alors possible d'envisager

des lexiques encore plus étendus (plusieurs centaines de milliers de mots).

L'utilisation de notre technique, en association avec des modèles de langages plus ou moins contraignants, conduira à la réalisation de systèmes de reconnaissance complets de parole continue ou de phrases énoncées en séparant les mots par de légères pauses.

Les matériels dont nous disposons actuellement (station de travail HP 9000 série 730 à 76 MIPS) effectuent la reconnaissance d'une cohorte en 3 minutes pour un lexique de 22.000 entrées. L'utilisation de cartes comportant quelques transputers permet de réaliser en temps réel la plupart des opérations coûteuses que l'on peut aisément paralléliser (Martin, 1991).

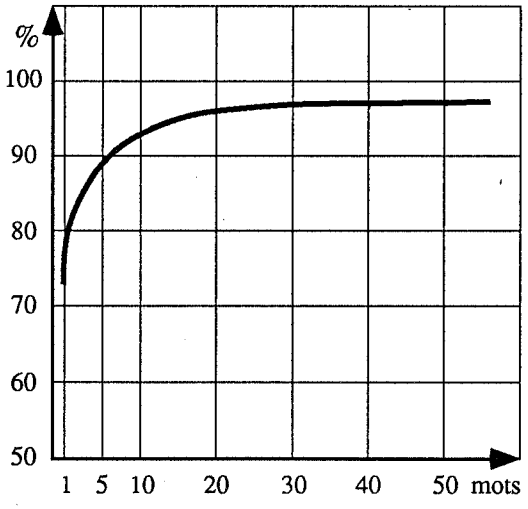


Figure 3 : Relation entre le nombre de mots de la cohorte et le taux de reconnaissance (présence du mot effectivement énoncé dans la cohorte d'hypothèses).

BIBLIOGRAPHIE

J. K. Baker (1989), "A Second-Generation Large Vocabulary System", *Speech Technology*, pp. 20-25.

J.F. Bonastre, H. Méloni, P. Langlais (1991), "Analytical strategy for speaker identification", *Proc. 2nd European Conference on Speech Communication and Technology*, Vol. 2, pp. 435-438, 24-26 septembre, Genova, Italy.

H. Bonneau, J.L. Gauvain (1987), "Vector Quantization for speaker adaptation", *Proc. IEEE ICASSP-87*, Vol. 3, p.1434.

R. Bulot (1987), "Techniques d'Intelligence Artificielle pour la reconnaissance de la parole : application au décodage acoustico-phonétique", *thèse de l'université d'Aix-Marseille II.*

Calliope (1989), *La parole et son traitement automatique*, Masson, pp. 512-543.

K. Choukri, G. Chollet, Y. Grenier (1986), "Spectral transformations through Canonical Analysis for speaker adaptation in ASR", *Proc. IEEE ICASSP-86*, Vol. 4, p. 2659.

J.L. Gauvain (1986), "A Syllable-Based Isolated Word Recognition Experiment", *Proc. IEEE ICASSP-86*, Vol. 3, pp. 35.10.1-35.10.4.

S.E. Levinson (1986), "Continuously Variable Duration Hidden Markow Models for Automatic Speech Recognition", *Computer Speech and Language*, Volume 1 (1), pp. 29-45.

C. Martin (1991), "Parallélisation au moyen de transputers d'algorithmes pour le décodage acoustico-phonétique de la parole", *Mémoire de DEA Informatique Signaux et Systèmes Option Parallélisme, Université de Nice-Sophia Antipolis.*

H. Méloni (1983), "Traitement des contraintes linguistiques en reconnaissance de la parole", *Revue Techniques et Sciences Informatiques*, Vol. 2, n° 5, pp. 349-363.

H. Méloni, R. Bulot (1986), "Un système de traitement de connaissances pour le décodage acoustico-phonétique", *Proc. ICA Symposium on speech recognition*, Montréal, 21-22 july 1986, pp. 24-26.

H. Méloni, P. Gilles (1991a), "Décodage acoustico-phonétique ascendant", *Revue Traitement du Signal*, Vol. 8, n° 2, pp. 107-114.

H. Méloni, P. Gilles, A. Betari (1991b), "Representation of acoustic and phonetic knowledge for speaker-independent recognition of small vocabularies", *Speech Communication*, Vol. 10, n° 2, pp. 145-154.

H. Méloni, P. Gilles (1991c), "Représentation de connaissances indépendantes du locuteur pour la reconnaissance de mots acoustiquement proches" ; *XIIème Congrès International des Sciences Phonétiques*, 19-24 Août 1991, Aix-en-Provence.

G. Pérennou, M. De Calmes (1986), "BDLEX : une base de données et de connaissances du français parlé", *Actes du séminaire GRECO-GALF Lexique et traitement automatique des langages*, Toulouse.

G. Shichman (1986), "An IBM PC Based Large vocabulary Isolated-Utterance Speech Recognizer" ; *Proc. IEEE ICASSP-86*, pp. 53-56.

T. Spriet (1990), "A Speech Understanding System" ; *Proc. International Conference on Spoken Language Processing*, 18-22 novembre, Kobe, Japan.

19es J.E.P. - Bruxelles, 19 au 22 mai 1992.

Apprentissage de modèles de Markov à l'aide de données réelles d'exploitation sélectionnées automatiquement.

Dominique MORIN
"Prosodie Informatique", 3 Villa Poirier, 75015 PARIS
CNET, LAA/TSS/RCP, BP 40, 22301 Lannion, France

Résumé

Cette communication concerne l'amélioration de systèmes de reconnaissance automatique de la parole, indépendants du locuteur. De précédents travaux ont montré qu'utiliser pendant l'apprentissage des données d'exploitation (enregistrées sur un serveur vocal en activité) validés manuellement, permet une diminution d'environ 30% du nombre des erreurs de reconnaissance en site réel d'exploitation. Nous utilisons les logiciels du CNET pour une reconnaissance de la parole basée sur une modélisation markovienne des mots (Hidden Markov Models).
Les travaux présentés ici ont pour objectif d'éviter la phase de validation manuelle des données d'exploitation. Plusieurs techniques de sélection vont être testées pour atteindre finalement une méthode entièrement automatique. Cette méthode, qui consiste à utiliser des modèles "poubelles", atteint les mêmes performances de reconnaissance que celles observées avec la validation manuelle.

1. INTRODUCTION

Les travaux présentés ici concernent l'amélioration des performances de systèmes de Reconnaissance Automatique de la Parole (R.A.P.), indépendants du locuteur, à travers le réseau téléphonique, à l'aide de "**données d'exploitation**" directement enregistrées sur un Serveur Vocal Interactif (S.V.I.) grand public en activité.
Bien que la R.A.P. atteigne des performances a priori satisfaisantes pour quitter les laboratoires, sa commercialisation en milieu industriel est encore rare. Les taux d'erreur de reconnaissance observés sur des systèmes interactifs grand public, utilisant la R.A.P., sont encore jugés trop élevés. De plus, l'ergonomie de ces systèmes est souvent problématique mais ceci n'est pas l'objet de cette étude.

En site réel d'exploitation, le taux d'erreur relevé est 2 à 4 fois supérieur à celui mesuré en laboratoire : les données utilisées en laboratoire et celles observées en exploitation sont donc différentes.
Le système de reconnaissance utilisé pour cette étude est basé sur une modélisation markovienne du vocabulaire de l'application. Ce type de reconnaissance nécessite une phase d'apprentissage qui permet, à l'aide d'un ensemble de données d'apprentissage, d'optimiser les paramètres des modèles. L'apprentissage n'est véritablement efficace que si les données d'apprentissage et les données à reconnaître sont de mêmes types: conditions d'enregistrement, lignes de transmission, et type des locuteurs. Classiquement, les données, dites **de laboratoire,** utilisées pour l'apprentissage, sont enregistrées sur le réseau téléphonique par des locuteurs à qui l'on demande de prononcer chacun des mots du vocabulaire de l'application. Ces données, ni très spontanées, ni très naturelles, sont peut-être trop différentes de celles d'exploitation. Cette base de données d'apprentissage a peut-être sa part de responsablilité dans la dégradation des performances de reconnaissance observée en exploitation.
Comme il est relativement facile de recueillir les données dites **d'exploitation**, observées en exploitation, a priori plus spontanées, nous avons testé leur utilisation en complément de celles de laboratoire pendant la phase d'apprentissage.
Une précédente étude a permis de tester l'influence optimale de ce nouveau type de données. Pour cela, nous avons étiqueté et validé manuellement une base de données d'exploitation recueillies sur un S.V.I. en situation réelle d'exploitation. L'étiquetage consiste à identifier chaque enregistrement à l'aide du nom du mot du vocabulaire de l'application qui a été prononcé. Quant à la validation, elle consiste à classer les enregistrements en CORRECTS ou

INCORRECTS. Nous avons ainsi construit une base de données d'exploitation contenant uniquement les enregistrements marqués CORRECTS afin de tester l'influence "**optimale**" de ces données. Et nous avons observé qu'en utilisant cette base de données en complément (puisque notre base de données d'exploitation n'est pas sufisament importante) de celle de laboratoire pendant l'apprentissage, les modèles résultants, appelés "modèles mixtes" entraînaient une diminution d'environ 30% du nombre des erreurs en situation réelle d'exploitation[1].
Il s'agit maintenant d'éviter les phases d'étiquetage et de validation manuelles. Ce sont les deux objectifs des travaux présentés dans cet article.
Le deuxième chapitre de cet article décrit le S.V.I. et les bases de données utilisées. La méthodologie suivie dans ces travaux est présentée dans le troisième. Les quatrième et cinquième chapitre décrivent les méthodes employées pour l'identification automatique et pour la validation automatique respectivement. Enfin, un sixième chapitre est consacré aux résultats de reconnaissance obtenus avec les modèles mixtes construits après identification et validation automatique.

2. LES BASES DE DONNEES

2.1. DESCRIPTION DU SERVEUR VOCAL

Le S.V.I. utilisé est le serveur MAIRIEVOX [2] développé au CNET de Lannion en 1988. Le système de R.A.P. utilisé détecte la parole à l'aide d'un automate bruit/parole. Actuellement, un enregistrement, constitué d'une séquence de trames, est précédé et suivi de plusieurs trames de silence pour éviter de devoir repérer exactement le début et la fin de parole. Les phases d'apprentissage et de reconnaissance sont réalisées par le logiciel PHIL86 [3] du CNET qui est basé sur une modélisation markovienne du vocabulaire (HMM). L'unité de base de modélisation peut être soit le mot (comme dans notre cas), soit une unité plus petite comme la syllabe, le phonème ou l'allophone [4]. Les procédures d'apprentissage et de reconnaissance utilisent l'algorithme de Viterbi.
Ce serveur MAIRIEVOX est un service téléphonique d'informations locales pour la ville de Lannion et ses environs. Il utilise un dialogue par menus, capable de reconnaître 21 mots isolés avec 6 mots uniquement autorisés dans chaque menu.

2.2. DESCRIPTION DES BASES DE DONNEES

Dans nos expériences, la parole utilisée est enregistrée à travers le réseau téléphonique français. Le signal analogique est numérisé en utilisant un Cofidec, loi A, à 8 khz. Après une préaccentuation du signal et un fenêtrage de Hanning de 32 ms avec recouvrement de 16 ms, 6 coefficients cepstraux sont calculés en utilisant 24 filtres répartis selon une échelle de MEL. Un 7ème coefficient représente le logarithme de l'énergie totale de la trame, et un 8ème la variation d'énergie entre les trames précédente et suivante. Nous allons dans ce paragraphe décrire les corpus de parole de laboratoire et d'exploitation.

2.2.1. BASE DE DONNEES DE LABORATOIRE

Pour construire cette base de données, nous avons enregistré un groupe de locuteurs à qui l'on a demandé de prononcer chacun des mots du vocabulaire de l'application, le plus naturellement possible, à travers le réseau téléphonique.
Après audition de la totalité de ce corpus, seuls les enregistrements jugés CORRECTS ont été retenus et portent le nom du mot qui a été prononcé.

2.2.2. BASE DE DONNEES D'EXPLOITATION

Pendant l'exploitation du serveur MAIRIEVOX, environ 20 000 détections ont été enregistrées [5]. Comme les utilisateurs ne suivent pas obligatoirement les instructions, certaines détections correspondent à des mots non autorisés ou de la parole continue ou des bruits. Nous avons établi 2 classes de détections:

- les CORRECTES qui contiennent uniquement un mot du vocabulaire de l'application; celui_ci étant correctement prononcé.
- les INCORRECTES comprenant :
  - du signal de parole correspondant à un mot du vocabulaire précédé ou suivi d'un mot hors vocabulaire ou bien uniquement un ou plusieurs mots hors vocabulaire: ces enregistrements sont marqués MAUVAIS,
  - uniquement du bruit (environnement ambiant du locuteur très bruyant, bruit de bouche, cris d'enfant, claquement de porte etc) et marqués BRUIT.

Environ 50% des enregistrements ont été annotés INCORRECTS.
Mais ce corpus d'exploitation, pas assez important, contient certains mots peu souvent prononcés. Afin d'éviter les conséquences de ce déséquilibre, en complément des tests sur l'ensemble des 21 mots du vocabulaire, nous avons réalisé les expériences sur un vocabulaire restreint aux 9 mots les plus fréquents de cette base de données.

2.3. DECOUPAGE ENTRE APPRENTISSAGE ET TEST.

Chaque base de données est divisée en deux parties de taille comparables: une pour l'apprentissage, et l'autre pour les tests de reconnaissance. Les bases de données d'apprentisage et de test de laboratoire sont prononcées par environ 250 locuteurs différents. Les bases de données d'exploitation comprennent environ 750 appels où les locuteurs ne sont pas forcément différents. La composition des corpus d'apprentissage et de test est donnée dans les tableaux 1 et 2 pour les vocabulaires de 21 mots et de 9 mots respectivement.

CORPUS	apprentissage		test	
	correct	incorrect	correct	incorrect
laboratoire	4893	0	4904	0
exploitation	4747	7238	4784	6774

Tableau 1: Composition des bases de données pour le vocabulaire de 21 mots.

CORPUS	apprentissage		test	
	correct	incorrect	correct	incorrect
laboratoire	2103	0	2110	0
exploitation	3694	2211	3744	1989

Tableau 2: Composition des bases de données pour le vocabulaire de 9 mots.

3. METHODOLOGIE

Une étude précédente a montré que l'utilisation de données d'exploitation CORRECTES pendant la phase d'apprentissage, en complément des données de laboratoire, améliore les performances de reconnaissance en situation réelle d'exploitation. Pour éviter les phases manuelles de validation et d'identification, il faut savoir détecter automatiquement ces données d'exploitation marquées manuellement CORRECTES, et les identifier. La méthodologie de construction automatique de la base de données d'exploitation qui participera à l'apprentissage des modèles mixtes est donc la suivante:

- on valide les données d'exploitation:
  On appelle "données **valides**", les données qui participeront à l'apprentissage. **L'objectif de la validation automatique est donc que les données valides correspondent à celles manuellement étiquetées CORRECTES.**
- on identifie les mots validés en utilisant le nom du mot reconnu par le système de reconnaissance. **L'identification optimale correspond donc aux taux d'erreurs de reconnaissance minimum des données valides.**
- on procède à l'apprentissage du modèle mixte, modélisant les 21 mots, en ajoutant, à la base de données de laboratoire, les données d'exploitation validées.

Nous allons détailler, dans les deux paragraphes qui suivent, les techniques de validation automatique, puis celles d'identification. Une dernière partie présentera les résultats de reconnaissance

4.VALIDATION AUTOMATIQUE

4.1 INTRODUCTION

Rappelons que l'objectif des travaux est d'atteindre, en utilisant ces données d'exploitation pendant l'apprentissage, les performances de reconnaissance obtenues avec la validation manuelle. Sélectionner la totalité des données CORRECTES, c'est à dire ne rejeter que les données INCORRECTES ou mal identifiées, n'est peut-être pas nécessaire (le rejet de quelques données CORRECTES et l'acceptation de quelques INCORRECTES peuvent être tolérables).
Pour ce faire, nous avons à notre disposition, un modèle laboratoire et notre base de données d'exploitation. Nous allons donc utiliser, pour la validation automatique:

- soit les scores de reconnaissance des données d'exploitation, récupérés sur le modèle de laboratoire,
- soit des modèles "poubelles",
- soit la combinaison de ces deux techniques.

Dans ces deux derniers cas, les modèles "poubelles" doivent au préalable être correctement appris.

Nous allons détailler ces trois méthodes dans les paragraphes qui suivent.

4.2 REJET PAR BALAYAGE DE SEUILS

Nous allons ici analyser le rejet par balayage de seuils sur les scores de reconnaissance Viterbi du HMM gagnant. Le score de reconnaissance (relatif à la probabilité d'émission) d'une donnée marquée CORRECTE est normalement supérieur à celui d'une donnée annotée INCORRECTE ... mais ceci n'est malheureusement pas toujours respecté. Cette méthode n'est donc pas très performante pour un rejet systématique des entrées INCORRECTES. Elle propose des taux non négligeables de fausses acceptation (F.A.) des données INCORRECTES et de faux rejet (F.R.) des CORRECTES: peut-être est-il possible de trouver le bon compromis entre ces deux taux.
Dans un premier temps, nous allons déterminer quel va être le score utilisé.

4.2.1. CHOIX DU SCORE A UTILISER

Le mot reconnu avec l'algorithme PHIL86 correspond au HMM proposant le meilleur score de Viterbi. Ce score est calculé sur toutes les trames de l'enregistrement y compris celles observées par les modèles de silence mis avant et après les modèles de mot. Il est d'autre part possible d'utiliser calculé sans tenir compte du score associé aux modèles de silences début et fin.

Nous allons, dans ce paragraphe comparer le comportement du rejet en utilisant ces deux types de scores.

Pour un modèle donné, nous avons tracé les 2 courbes de rejet (cf. figure1) par balayage de seuils sur chacun des 2 types de scores. On trouve, en abscisse, le taux d'erreur de reconnaissance des données manuellement étiquetées CORRECTES et en ordonnée, le taux de Fausses Acceptations des données manuellement étiquetées INCORRECTES.

La courbe de rejet idéale serait celle qui serait confondue avec les axes et qui correspondrait à 0% de F.A. des données INCORRECTES et 0% d'erreur de reconnaissance des données CORRECTES. Le figure 1 montre que la courbe la plus près de ce cas idéal correspond à celle où le score Viterbi est calculé sans tenir compte du score associé aux modèles de silences de début et de fin de mot. Et ce résultat est sans équivoque : les courbes n'ont aucun point d'intersection.

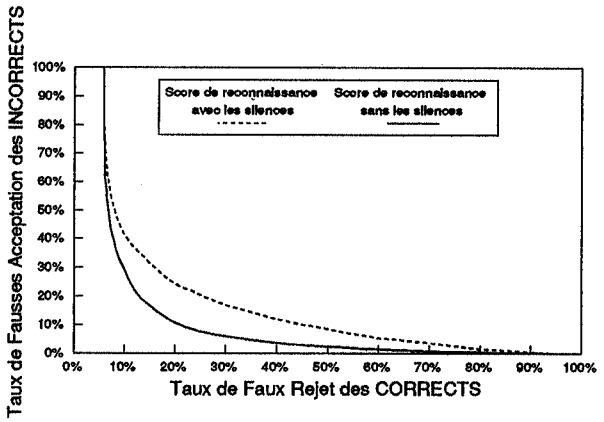


Figure 1: Comparaison du comportement du rejet par balayage de seuils de rejet sur les différents scores de reconnaissance Viterbi.

4.2.1. CALCUL DU SEUIL DE REJET

Nous avons vu précédemment qu'en balayant une plage de seuils arbitraires, le rejet sur le score sans les silences de début et de fin de mot donnait les meilleures performances. Il faut maintenant déterminer, de façon automatique, la valeur du seuil de rejet à utiliser.

Pour cela, nous avons à notre disposition une base de données de laboratoire étiquetée et validée manuellement, un modèle dit "de laboratoire" appris avec cette base de données et une base de données d'exploitation identifiée mais pas validée. Nous avons récupéré, en utilisant le modèle laboratoire, les scores de reconnaissance de chaque enregistrement d'exploitation en les classant. Nous avons ensuite pris, comme seuil de rejet, les scores correspondant à différents taux de rejet de la base de données d'exploitation. Les meilleurs résultats de reconnaissance ont été obtenus avec les seuils correspondant à des taux de rejet compris entre 40% et 60%. Mais la valeur de ces taux n'est pas surprenante puisque notre base de données d'exploitation contient environ 50% de données manuellement marquées INCORRECTES et donc à rejeter.

Cette méthode de calcul de seuil de rejet n'est pas entièrement satisfaisante puisqu'elle dépend de la composition de la base de données d'exploitation (i.e. pourcentage de données correctes)

Ces techniques de rejet par seuillage de score de reconnaissance nous a permis d'atteindre les performances de reconnaissance espérées mais nous n'avons trouvé aucune méthode satisfaisante de calcul automatique du seuil de rejet optimal. Nous avons donc opté pour l'utilisation de modèles "poubelles".

4.3. UTILISATION DE MODELES "POUBELLES"

Une autre technique de rejet consiste à utiliser des modèles "poubelles" qui auront pour but de "piéger" les données à rejeter. Le problème est alors de déterminer les données qui doivent être utilisées pour l'apprentissage de ce type de modèles.

Nous avons ainsi créé les modèles de vocabulaire suivants:

a) un premier incluant 3 modèles "poubelles" appris l'un avec les données étiquetées manuellement BRUIT et les deux autres avec celles étiquetées manuellement MAUVAIS: puisque l'on utilise l'étiquetage manuel, on ne pourra pas considérer cette méthode comme étant automatique.

b) un second où les modèles "poubelles" ont été construits avec les données d'exploitation proposant les scores de reconnaissance, en utilisant le modèle laboratoire, inférieurs à un seuil de rejet. Cette technique, qui utilise un seuil de rejet sera qualifiée de quasi-

automatique puisque nous n'avons pas trouver de méthode de calcul de seuil automatique.

c) un dernier contenant 2 modèles "poubelles" appris chacun avec des vocabulaires différents de celui de l'application: les 10 chiffres français (alors que le vocabulaire de l'application ne contient pas de chiffre) et les 500 mots les plus fréquents du français. La validation à l'aide de ce modèle est, cette fois-ci, entièrement automatique.

5. IDENTIFICATION AUTOMATIQUE

Afin de pouvoir utiliser la base de données d'exploitation pendant l'apprentissage, il est nécessaire d'identifier chaque enregistrement à l'aide du nom du mot du vocabulaire prononcé. Sans intervention manuelle, l'identification va donc consister à attribuer à chaque enregistrement le nom du mot du vocabulaire reconnu. Les erreurs d'identification automatique correspondent donc aux erreurs de reconnaissance.

Le mot reconnu avec l'algorithme PHIL86 correspond au HMM proposant le meilleur score de Viterbi. Ce score est calculé sur toutes les trames de l'enregistrement y compris celles observées par les modèles de silence mis avant et après les modèles de mot.

Un test de reconnaissance, effectué sans tenir compte du score associé aux modèles de silences début et fin, n'a pas amélioré les performances.

6. TESTS DE RECONNAISSANCE

Dans ce paragraphe, nous allons analyser les résultats de reconnaissance obtenus en utilisant des modèles mixtes appris avec des données d'exploitation identifiées et validées avec les méthodes décrites précédemment. Rappelons qu'un modèle mixte utilise, pendant sa phase d'apprentissage, des données d'exploitation en complément de celles de laboratoire.

Nous avons ainsi construit 4 modèles mixtes:

- "seuil" où la base de données d'exploitation a été construite en utilisant la méthode de rejet par seuillage du score de reconnaissance Viterbi décrite au §4.2.1. Le seuil de rejet utilisé ici est le score de reconnaissance qui correspond au rejet d'environ 50% des données d'exploitation en utilisant le modèle de laboratoire.
- "pb_man" où les données d'exploitation ont été sélectionnées en utilisant 3 modèles "poubelles" appris avec les données marquées manuellement MAUVAIS ou BRUIT (cf..§4.3.a).
- "pb_semi_auto" où la validation est effectuée avec 3 modèles "poubelles". Cette fois-ci, ces modèles ont été appris avec les données d'exploitation validées avec la méthode énoncée au §4.3.b. Le problème de calcul du seuil s'est encore posé d'où le terme semi-automatique.
- "pb_auto" où la base de données d'exploitation a été sélectionnée en utilisant un vocabulaire contenant 2 modèles "poubelles" appris avec 2 vocabulaires différents de celui de l'application comme nous l'avons déjà décrit au §4.3.c.

Les résultats de reconnaissance de ces modèles apparaissent sur les figures 2 et 3 pour les vocabulaires de 21 mots et de 9 mots respectivement En abscisse, on trouve le nom du modèle mixte, en ordonnée, le taux d'erreur de reconnaissance des données d'exploitation marquées manuellement CORRECTES. Ces 2 figures indiquent d'autre part, sous forme d'histogramme, la composition des bases de données d'exploitation qui ont participé à l'apprentissage: on peut voir le taux de F.A. des données étiquetées INCORRECTES et le taux de F.R des données étiquetées manuellement CORRECTES.

Les taux de reconnaissance obtenus à l'aide des 4 modèles sont à comparer à ceux obtenus avec les modèles laboratoire et mixte de référence.

En se référant aux figures 2 et 3, on remarque que les **performances obtenues avec la validation manuelle** (celles relevées sur le modèle mixte) **sont atteintes avec nos nouveaux modèles mixtes**. Nous avons donc atteint notre objectif puisque, pb_auto a utilisé une validation complètement automatique.

On peut remarquer qu'en ce qui concerne la composition des bases de données d'exploitation de ces 4 modèles, les taux de données INCORRECTES parmi les données validées varient de 10% à 20% quand les taux données CORRECTES validées, parmi les CORRECTES, varient de 70% à 85% pour le vocabulaire des 21 mots. Pour le vocabulaire des 9 mots, les taux de données INCORRECTES validées oscillent entre 10% et 25% quand les taux de CORRECTES validées varient de 80% à 85%. D'autre tests ont montrés que des taux de données INCORRECTES validées supérieurs perturbent la modélisation.

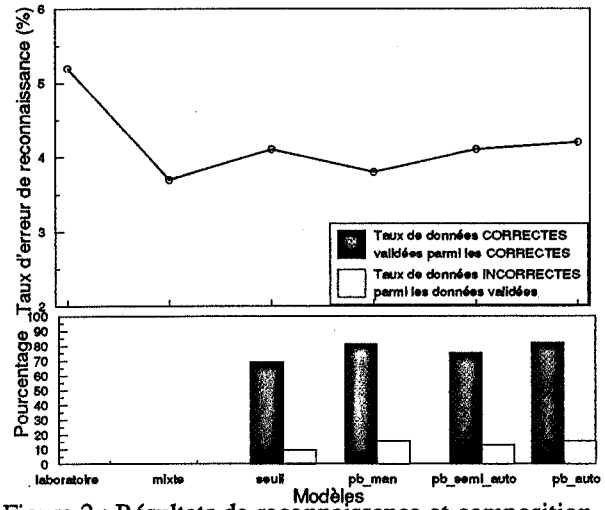


Figure 2 : Résultats de reconnaissance et composition des bases de données d'exploitation avec le vocabulaire de 21 mots

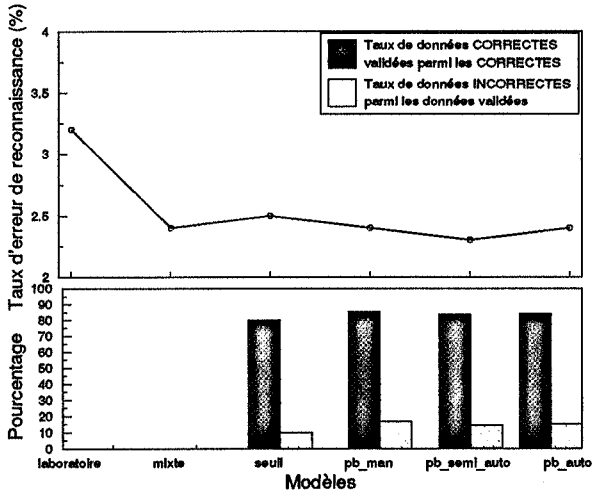


Figure 3 : Résultats de reconnaissance et composition des bases de données d'exploitation avec le vocabulaire de 9 mots

5.CONCLUSIONS

Alors que nous avions montré l'influence positive des données réelles d'exploitation (diminution de 30% du nombre des erreurs) manuellement marquées CORRECTES dans l'apprentissage de modèles de Markov, nous venons d'atteindre, dans cette étude, les mêmes performances en utilisant des procédures automatiques.

L'objectif était de valider puis d'identifier automatiquement une base de données d'exploitation contenant suffisamment de données CORRECTES. Plusieurs méthodes manuelles ou semi-automatique ont été testées avec succès. Et, finalement, une technique entièrement automatique a atteint les performances espérées.

Pour la phase de validation, cette méthode consiste à utiliser des modèles "poubelles" pour piéger les données exploitation INCORRECTES à rejeter de la base de données d'apprentissage. L'automatisation complète vient du fait que les modèles "poubelles" sont appris avec **des vocabulaires différents de ceux de l'application.**

Ces données d'exploitation ainsi validées puis identifiées, en participant à l'apprentissage de modèles mixtes en complément des données de laboratoire, ont permis une diminution d'environ **30% du nombre des erreurs par rapport à l'utilisation du modèle laboratoire.** L'objectif est donc atteint bien que la base de données d'exploitation validée contienne environ 15% de données INCORRECTES. Bien entendu, il est nécessaire de vérifier ces résultats en utilisant une base de données d'exploitation plus importante d'une part et d'autres vocabulaires d'autre part.

Remerciements

Je tiens à remercier Jean Monné et Denis Jouvet pour leur aide et leurs suggestions.

Bibliographie

[1] D. Morin : "Influence of field data in HMM training for a vocal server" : Eurospeech'91, Gênes, septembre 91, Vol. 2, 735-738.

[2] C. Gagnoulet, D. Jouvet et J. Damay: "MAIRIEVOX : A Voice-Activated Information System" : Speech Communication 10, 1991, 23-31.

[3] D. Jouvet, J. Monné, D. Dubois : "A New-Network-Based Speaker Independant Connected-Word Recognition System" : Proc. IEE Int. Conf. ASSP 1986, Tokyo, p 1109-1112.

[4] D. Jouvet, K. Bartkova, J. Monné "On the modelization of allophones in an HMM based speech recognition system" : Eurospeech'91, Gênes, septembre 1991, Vol. 2, 923-926.

[5] L. Mathan et D. Morin : "Speech field databases: developpment and analysis", Eurospeech 91, Gênes, septembre 1991, Vol. 2, 509-512.

19es J.E.P. - Bruxelles, 19 au 22 mai 1992.

RECHERCHE DES N MEILLEURES SOLUTIONS EN RECONNAISSANCE DE MOTS CONNECTES

Mohamed Nabil LOKBANI

Centre National d'Etudes des Télécommunications
LAA/TSS/RCP - Route de Trégastel
22301 LANNION - FRANCE

Résumé

Cet article présente une méthode de recherche des N meilleures solutions en reconnaissance de la parole. Cette recherche est introduite essentiellement pour permettre un post-traitement (contraintes syntaxiques, discriminations,...). La méthode employée, consiste à utiliser l'algorithme de Viterbi dans la phase aller (effectuée de manière synchrone) et l'algorithme A* dans la phase retour (effectuée de manière asynchrone). L'algorithme de Viterbi se charge de calculer et de mémoriser les différentes probabilités d'émission (meilleur chemin) à chaque instant t du signal à reconnaître. Apres cette phase aller, l'algorithme A* est utilisé pour remonter le chemin à partir du dernier noeud ; l'estimation de la portion du signal non encore traitée provient des valeurs mémorisées pendant la phase aller. Cette estimation est optimale et conduit alors à une recherche très efficace des N meilleures solutions. Les différents résultats présentés à la fin de cet article montrent l'évolution du taux d'erreur résiduel en fonction des N meilleures solutions.

1- INTRODUCTION

Les systèmes de reconnaissance de la parole développés Au CNET, reposent sur une modélisation statistique du vocabulaire de chaque application. Ces systèmes de reconnaissance, qui utilisent un réseau compilé (une procédure d'optimisation du réseau est introduite dans cette étape), ont été essentiellement développés pour le traitement de petits vocabulaires (mots isolés ou mots connectés) et servent pour la réalisation de serveurs vocaux interactifs (SVI).

L'utilisation d'un réseau compilé facilite l'introduction de règles syntaxiques, lexicales, phonologiques et la mise en oeuvre de modélisations contextuelles (allophones). La taille du réseau compilé rend le système inadapté pour le traitement de syntaxes complexes. Pour y remédier, nous proposons une syntaxe plus "libre" où toutes les successions de mots (ou expressions) sont possibles. La meilleure solution (au sens des probabilités) obtenue par ce système n'est alors pas forcément correcte. Pour cela, nous développons une méthode de recherche des N meilleures solutions.

La rechèrche des N meilleures solutions ne présente un intérêt que par la mise en oeuvre d'un post-traitement. Les post-traitements envisagés actuellement ont pour objectif :

* de rechercher dans ces N meilleures solutions, la première solution qui respecte la véritable syntaxe de l'application.

* de réaliser un post-traitement segmental sur les N solutions proposées.

Le premier paragraphe présente le système de reconnaissance de la parole développé au CNET. Par la suite, on décrit la manière de rechercher les N meilleures solutions. Avant de conclure, on présente un ensemble de résultats obtenus sur les bases du laboratoire.

2- LE SYSTEME DE RECONNAISSANCE DE LA PAROLE

Une des techniques utilisées dans les systèmes de reconnaissance, repose sur l'utilisation des modèles de Markov cachés [1]. Toute la connaissance, syntaxique, lexicale et acoustique, peut être explicitement décrite dans un réseau modélisant l'ensemble des phrases admises par l'application [2]. Dans ce cas la connaissance est étroitement intégrée dans le système. La phase de reconnaissance consiste alors à parcourir le réseau de façon homogène pour y trouver le mot (ou la phrase) reconnu. Ce parcours peut être réalisé par l'utilisation de l'algorithme A*, méthode de recherche dans un graphe ; ou bien par l'algorithme de Viterbi, adapté aux réseaux compilés,

et appliqué de manière synchrone (i.e. trame après trame). C'est l'approche actuellement utilisée au CNET pour la reconnaissance de mots isolés ou connectés.

2.1- Description du système de reconnaissance

Le système de reconnaissance développé au CNET, permet la reconnaissance de vocabulaire dont la taille reste limitée à une centaine de mots environ, de manière indépendante du locuteur. Il autorise la reconnaissance de mots isolés ou de mots connectés. Il permet l'introduction dans les modèles de règles syntaxiques, lexicales et phonologiques. Il utilise l'approche par modèles de Markov cachés pour représenter l'ensemble des mots du vocabulaire. La méthode utilisée pour la phase de reconnaissance est la méthode statistique. Ainsi pour la reconnaissance d'un mot, on calcule la probabilité d'émission de ce mot sur chacun des modèles statistiques. Le mot reconnu est celui correspondant au modèle qui a donné la plus grande probabilité, calculée par l'algorithme de Viterbi.

2.2- Modélisation statistique

Les modèles de Markov sont représentés par un ensemble d'états q_i, de transitions $a_{i,j}$ et de fonctions de densité de probabilité $B_{i,j}$. Ces fonctions de densité de probabilité sont associées aux transitions, et sont choisies continues et gaussiennes. Chaque fonction de densité de probabilité possède deux paramètres, le vecteur des moyennes $m_{i,j}$ et la matrice de covariance diagonale $\Sigma_{i,j}$.

L'analyse acoustique calcule toutes les 16 ms (fenêtre de Hanning de 32 ms avec un recouvrement de 50%) 8 coefficients cepstraux obtenus à partir de l'échelle Mel (MFCC), complétés par un paramètre d'énergie, plus les dérivées premières et secondes de ces 9 coefficients.

Au cours de la reconnaissance, on s'intéresse à calculer la probabilité maximale d'observation de l'ensemble des trames du mot (ou de la phrase) inconnu. L'algorithme de Viterbi est chargé de calculer $\{\Phi[\tau, q_i], \forall i\}$, à partir de $\{\Phi[\tau-1, q_i], \forall i\}$, où $\Phi[\tau, q_i]$ est la probabilité maximale d'observation des τ premières trames, le long des chemins atteignant l'état q_i au temps τ. Pour chaque trame et pour tous les états de la chaîne, nous utilisons la formule de récurrence suivante qui établit qu'un chemin de longueur τ résulte de la prolongation d'un chemin de longueur $\tau-1$ par une transition entre états et l'observation de la trame $X[\tau]$ au cours de cette transition.

$$\Phi[\tau, q_i] = \underset{q_j}{\text{Max}}\ \Phi[\tau-1, q_j].a_{ji}.B_{ji}(X[\tau])$$

Ainsi, en notant q_F le dernier état de la chaîne, $\Phi[T, q_F]$ est la probabilité cherchée.

2.3- Modélisation d'une application

Une syntaxe décrit l'ensemble des phrases admises par l'application (exemple : les chiffres de 0 à 9, précédés et suivis d'un silence). Une description lexicale transforme chaque mot du vocabulaire, en terme d'unités arbitraires plus petites ; des phonèmes par exemple (z, ei, r ,au pour le chiffre 0). Par la suite, chaque unité de la modélisation est représentée par un modèle de Markov caché, on parle alors de description acoustique.

Une application peut être modélisée de différentes manières :

* *modélisation par mots* : le niveau lexical n'intervient pas dans, de ce fait chaque mot est représenté par un modèle régulier à M états.

* *modélisation par phonèmes* : le niveau lexical est pris en compte. Chaque unité phonétique doit être modélisée au niveau acoustique.

* *modélisation par allophones* : la réalisation acoustique d'un phonème [3] étant très dépendante du contexte dans lequel il apparaît (c'est-à-dire des phonèmes qui l'entourent), il est naturel d'utiliser une modélisation qui prend en considération ces influences contextuelles. C'est l'approche par allophones qui consiste à utiliser pour chaque phonème un modèle acoustique comportant plusieurs entrées et sorties, choisies et validées en fonction du contexte dans lequel le phonème apparaît.

3- RECHERCHE DES N MEILLEURES SOLUTIONS

L'intérêt porté pour les N meilleures solutions, ces dernières années, est essentiellement dû à 2 points importants :

* *la recherche de solutions syntaxiquement différentes* : Cette approche permet de prendre en considération des contraintes syntaxiques n'apparaissant pas dans la modélisation. Dans le cas de réseaux compilés, on peut alors utiliser une syntaxe possédant plus de liberté. Par exemple, tout mot ou expression, peut suivre tout autre mot ou expression. De ce fait la compilation du réseau ne dupliquera pas de modèles, et le réseau obtenu sera beaucoup plus compact. Malheureusement, la meilleure solution au sens des probabilités n'est pas forcément syntaxiquement correcte, d'où l'intérêt de rechercher d'autre solutions syntaxiquement différentes.

* *un post-traitement segmental peut être appliqué* : dans la phase de recherche des N meilleures solutions, la mémorisation de l'ensemble de l'alignement peut être envisageable pour par-exemple permettre la prise en compte d'informations segmentales (durées,...) lors du post-traitement.

Plusieurs méthodes ont été proposées, pour rechercher les N meilleures solutions.

Myers et Rabiner [4] ont proposé à travers l'algorithme Level Building, de rechercher les N meilleures solutions comme étant une combinaison et une segmentation de la meilleure solution obtenue pour un niveau donné. Cependant les N meilleures solutions obtenues ne sont pas exactes.

Lee et Rabiner [5] ont utilisé la même idée par la suite sur l'algorithme One-Pass pour rechercher exactement la seconde meilleure solution.

Steinbiss [6] proposa cette fois ci de rechercher les N meilleures séquences de mots intermédiaires. La progression de recherche s'effectue entre noeuds du réseau acoustique. A chaque noeud du réseau sont associés les N meilleurs scores des différents noeuds pouvant lui être associés. Cette procédure synchrone permet d'obtenir les N meilleures solutions, mais le coût en calcul est très élevé.

Schwartz et Chow [7] ont proposé, de combiner les chemins identiques aboutissant au même état acoustique. Pour cela, ils ont utilisé un algorithme de Viterbi modifié avec une recherche en largeur. Si deux noeuds arrivent sur un même état au même instant et véhiculent la même séquence de mots, le score affecté au noeud cible est une combinaison des deux scores, sinon un nouveau chemin est crée dans la liste.

Paul [8, 9] a proposé lui d'utiliser l'algorithme A* dans la phase de décodage (phase aller). Pour cela, il devait estimer chaque fois la portion du signal non encore décodée. L'idée de base provient de Jelinek [10] , où il a développé cet algorithme comme étant une méthode de décodage séquentiel en transmission de l'information.

Austin et al [11], utilisent deux phases de calcul. La phase aller où ils calculent et mémorisent l'ensemble des probabilités par l'algorithme de Viterbi. Dans la phase retour, ils reprennent le même procédé de la phase aller mais à chaque instant, ils somment pour chaque mots trouvé, le score obtenu dans la phase aller avec le score calculé. Ce score total est sauvegardé dans une liste. La procédure est répétée jusqu'à obtenir le noeud de départ. A la fin, les N meilleures solutions sont récupérées de cette liste. Cette procédure dérive de l'algorithme Baum-Welch.

Soong et Huang [12], divisent également le procédé de recherche des N meilleures solutions en deux phases. Une phase aller (phase synchrone) pour calculer les différentes probabilités d'émission et une phase retour (phase asynchrone) pour remonter le chemin à partir du dernier noeud jusqu'au noeud de départ en utilisant l'algorithme A*. Cette méthode permet d'obtenir exactement les N meilleures solutions.

3.1 Algorithme A*

L'algorithme A* [13] est un algorithme permettant de trouver le meilleur chemin entre un noeud racine (ici l'état q_F) et un noeud objectif (ou cible, ici l'état q_0). Il utilise pour cela une fonction coût :

$$f(n_i) = g(n_i) + h(n_i)$$

où n_i est un noeud intermédiaire du réseau acoustique, $g(n_i)$ le coût (connu) de la racine jusqu'au noeud n_i et $h(n_i)$ le coût (estimé) du noeud n_i jusqu'au noeud objectif.

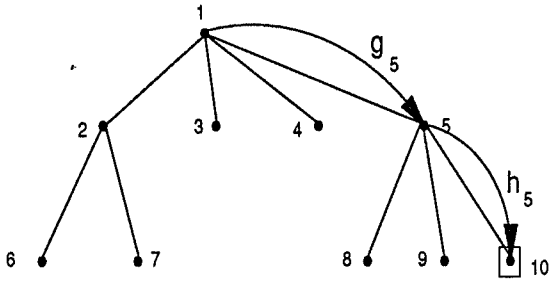


Figure 1 - Exemple de recherche du noeud final

La procédure de recherche du noeud final s'effectue de la manière suivante : on développe l'ensemble des noeuds partant de la racine. Si un noeud correspond au noeud final, on arrête la procédure de recherche sur le noeud trouvé, sinon, on calcule la fonction coût de chaque noeud. On introduit cette valeur dans une liste. La liste est triée par valeur croissante de la fonction coût f. Le noeud suivant à développer est celui se trouvant au sommet de cette liste. Ainsi de suite jusqu'à trouver le noeud final. Pour l'exemple présenté sur la figure 1, le noeud de départ est le noeud 1, le noeud d'arrivée est le noeud 10. La fonction coût pour le noeud 5, n'est autre que le coût de passage du noeud 1 à 5, ajouté à cela une estimation sur le chemin restant à parcourir (de 5 à 10).

Les propriétés de terminaison et d'admissibilité permettent au A* de trouver un chemin optimal entre le noeud de départ et le noeud final, si ce dernier existe. L'algorithme A* s'arrête soit sur le noeud final ou sur un échec (dans le cas ou il n'existe pas d'objectif à atteindre) et ceci dans un graphe fini.

L'efficacité de l'algorithme A* est liée à la manière dont la fonction h est estimée. Si h est un estimateur parfait alors l'algorithme A* convergera immédiatement vers l'objectif à atteindre. Dans le cas ou h est sur-estimée, on ne peut avoir l'assurance de trouver le chemin solution le moins coûteux. Par-contre, si h est sous-estimée, la meilleure solution peut être trouvée. Meilleur est h, plus nous obtiendrons rapidement la meilleure solution.

3.2 Méthode de recherche des N meilleures solutions

La méthode de recherche des N meilleures solutions va se dérouler en deux phases :

Phase aller : l'algorithme de Viterbi se charge de calculer et de mémoriser les différentes probabilités d'émission (meilleur chemin) à chaque instant t du signal à reconnaître. Cette procédure est la même que celle décrite auparavant (2.2) sauf que ici, on introduit la procédure de mémorisation des probabilités. C'est la phase aller de notre recherche (phase synchrone).

Phase retour : la phase retour (phase asynchrone) est réalisée par un algorithme de recherche dans un graphe, l'algorithme A*. Cet algorithme est utilisé pour remonter le chemin à partir du dernier noeud ; l'estimation de la portion du signal non encore comparée pour la fonction h, provient des valeurs mémorisées pendant la phase aller. La fonction h n'est plus estimée, elle est connue avec précision. Cette estimation optimale conduit alors à une recherche très efficace des N meilleures solutions. Cette efficacité nous a conduit à adopter cette méthode pour la recherche des N meilleures solutions.

3.3 Description de la fonction coût utilisée

La fonction coût f est déterminée par la somme de la fonction h dans la phase aller, et de la fonction g dans la phase retour. h est déterminée par

$$h[\tau, q_j] = \underset{q_i}{\text{Max}}\, h[\tau-1, q_i].a_{ij}.B_{ij}(X[\tau])$$

où $h[\tau, q_j]$ est la probabilité maximale d'observation des τ premières trames, le long des chemins atteignant l'état q_j au temps τ. Ce n'est autre que la fonction calculée en (2.2), quant à g, elle est déterminée par

$$g[\tau, q_j] = \underset{q_i}{\text{Max}}\, g[\tau+1, q_i].a_{ji}.B_{ji}(X[\tau+1])$$

où $g[\tau, q_j]$ est la probabilité maximale d'observation des (T-τ) dernières trames, le long des chemins partant de l'état q_j au temps τ, $g[\tau+1, q_i]$ est la probabilité maximale d'observation des (T-τ-1) dernières trames, le long des chemins partant de l'état q_i au temps $\tau+1$, a_{ji} la transition de passage de l'état q_j à l'état q_i et $B_{ji}(X[\tau+1])$ est la valeur de distribution associée à cette transition à l'instant $\tau+1$.

La fonction coût à l'instant τ, et pour l'état q_j, est :

$$f[\tau, q_j] = g[\tau, q_j] + h[\tau, q_j]$$

(Le graphe ci-dessous représente une partie de la progression de recherche des N meilleures solutions).

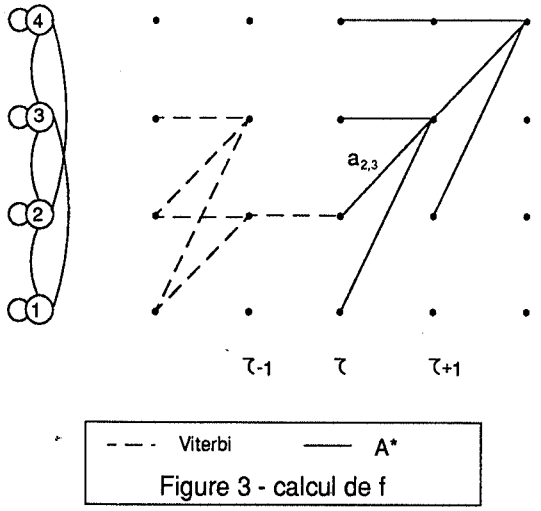


Figure 3 - calcul de f

Figure 2 - Progression de recherche des N meilleures solutions par Viterbi et A*

4- RESULTATS

4.1 Bases de données

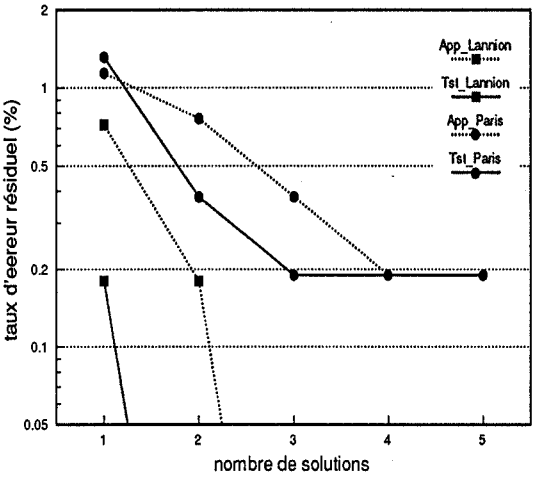
Trois bases de données ont été utilisées pour les différents tests. Ces bases ont été enregistrées à travers le réseau téléphonique par près de 800 locuteurs de différentes régions de France. Ces bases sont : la base des chiffres isolés (0 à 9), la base des nombre à deux chiffres (00 à 99, 25 nombres par locuteurs) et la base du **Trégor** (36 mots isolés). Cette dernière base a servi pour la réalisation du serveur vocal MAIRIEVOX [14] .

4.2 Tests et commentaires

Les résultats sont présentés sous forme de graphe, l'axe des x correspond au nombre de solution et l'axe des y correspond au taux d'erreur résiduel pour les N solutions désirées. Le nombre maximal de solutions est égal à 5. Pour des raisons pratiques, des trois bases citées plus haut, on a extrait les enregistrements correspondants à deux villes, Paris et Lannion pour les bases d'apprentissage (App) et de test (Tst). On a choisi une modélisation par mots. Le nombre d'états affecté à chaque mot est égal à 30.

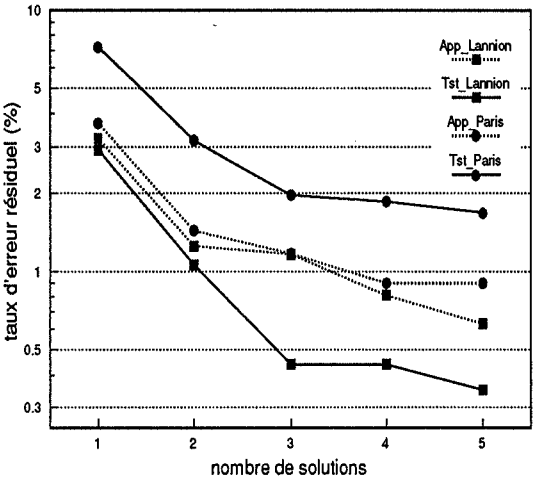
Le nombre de coefficients est égal à 27. Les modèles ont été appris sur environ 400 locuteurs et cela pour chacune des trois bases citées plus haut.

*** Base des chiffres isolés** : les chiffres de 0 à 9. Les bases d'apprentissage et de test pour la ville de Lannion comportent respectivement 560 et 559 enregistrements. Quant à la ville de Paris, elles comportent respectivement 522 et 535 enregistrements.

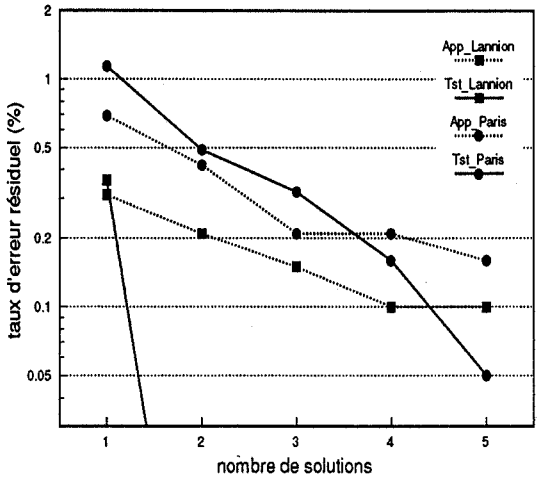


Pour les bases App et Tst de la ville de Lannion, le taux d'erreur résiduel atteint la valeur 0 pour respectivement N=3 et N=2.

*** Base des nombres** : les nombre à 2 chiffres (00 à 99). Les bases d'apprentissage et de test pour la ville de Lannion comportent respectivement 1117 et 1131 enregistrements. Quant à la ville de Paris, elles comportent respectivement 1114 et 1073 enregistrements.



*** Base Trégor** : 36 mots isolés. Les bases d'apprentissage et de test pour la ville de Lannion comportent respectivement 1936 et 1967 enregistrements. Quant à la ville de Paris, elles comportent respectivement 1884 et 1848 enregistrements.



Pour la base Tst de la ville de Lannion, le taux d'erreur résiduel atteint la valeur 0 pour N=2.

De ces résultats, on peut tirer plusieurs commentaires :

* Sur les 3 bases de données, la recherche de la 2ème solution, permet de réduire par 2 le taux d'erreur résiduel.

* Compte tenu des courbes ci-dessus, un post-traitement efficace devrait permettre une réduction substantielle du taux d'erreur.

5- CONCLUSION

La méthode proposée auparavant permet d'obtenir exactement les N meilleures solutions. Cependant la recherche des N solutions ne présente d'intérêt que si l'on effectue un post-traitement permettant de choisir entre les solutions proposées : prise en comte de contraintes syntaxiques, utilisation d'informations segmentales, etc...

Les tests effectués montrent qu'en vue d'un post-traitement segmental la recherche des 3 à 5 meilleures solutions semble suffisante, et permet d'espérer une réduction substantielle du taux d'erreur.

REMERCIEMENTS

Je tiens à remercier Denis Jouvet pour son aide et ses suggestions.

BIBLIOGRAPHIE

[1] F. Jelinek : "Continuous Speech Recognition by Statistical Methods" ; IEEE, Vol 64, Avril 1976.

[2] D. Jouvet : "Reconnaissance de Mots Connectés Indépendamment du Locuteur Par des Méthodes Statistiques" ; Thèse Doctorat de L'ENST, Juin 1988.

[3] D. Jouvet, K. Bartkova, J. Monné : "On The Modelization of Allophones in HMM Based Speech Recognition System" ; EUROSPEECH, Gènes, Septembre 1991.

[4] C. S. Myers, L. R. Rabiner : "Connected Digit Recognition Using a Level-Building DTW Algorithm" ; IEEE Trans On ASSP-29, No. 3, Juin 1981.

[5] C. LEE, L. R. Rabiner : "A Frame-Synchronous Network Search Algorithm for Connected Word Recognition" ; IEEE Trans On ASSP, Vol 37, No. 11, Novembre 1989.

[6] V. Steinbiss : "Sentence-Hypotheses Generation in Continuous-Speech Recognition System" ; EUROSPEECH, Paris, Septembre 1989.

[7] R. Schwartz, Y. Chow : "The N-Best Algorithm : An Efficient And Exact Procedure For Finding The N most Likely Sentence Hypotheses" ; ICASSP, Albuquerque, Avril 1990.

[8] D. B. Paul : "A CSR-NL Interface Specification Version 1.5" ; DARPA Speech and Natural Language Workshop, Octobre 1989.

[9] D. B. Paul : "Algorithms for an Optimal A* Search and Linearizing the Search in the Stack Decoder" ; ICASSP, Toronto, Mai 1991.

[10] F. Jelinek : "A Fast Sequentiel Decoding Algorithm Using a Stack" ; IBM J. Res. Develop, Vol 13, Novembre 1969.

[11] S. Austin, R. Schwartz et P. Placeway : "The Forward-Backword Search Algorithm" ; ICASSP, Toronto, Mai 1991.

[12] F.K.Soong, E. Huang : "A Tree-Trellis Based Fast Search for Finding the N Best Sentence Hypotheses in Continuous Speech Recognition" ; ICASSP, Toronto, Mai 1991.

[13] N. J. Nillson : "Principles of Artificial Intelligence" ; Tioga publishing, 1980.

[14] C. Gagnoulet, D. Jouvet : "Reconnaissance de la parole et modélisation statistique : expérience du CNET" ; Traitement du Signal, Volume 7, No 4, 1990.

19es J.E.P. - Bruxelles, 19 au 22 mai 1992.

Classification segmentale de voyelles avec des réseaux à unités gaussiennes ou sigmoïdales

Denys Boiteau

CNET Lannion LAA/TSS/RCP, 22301 Lannion FRANCE

Résumé

Cet article présente les expériences que nous avons réalisées avec des réseaux à unités sigmoïdales (MLP pour Multilayer Perceptron) et gaussiennes (RBF pour Radial Basis Functions) sur une tâche de classification segmentale de phonèmes. Cette approche segmentale de la reconnaissance de la parole permet l'intégration de paramètres temporels inaccessibles dans une approche par trames. Les expériences rapportées ici montrent leur intérêt pour une meilleure discrimination phonétique.

De plus, les expériences montrent que, pour obtenir les meilleures performances, le choix des coefficients d'entrée, corrélés ou non, dépend de la partition de l'espace d'entrée réalisée par les cellules cachées du réseau. Au contraire des MLP, les réseaux à cellules gaussiennes à matrices de covariance diagonales obtiennent leurs meilleures performances avec des coefficients peu ou pas corrélés en entrée.

1. INTRODUCTION

Les réseaux connexionnistes sont largement employés pour de nombreuses tâches: classification, prédiction ou compression de données. Ces réseaux se sont développés avec la mise au point de l'algorithme de rétropropagation du gradient pour l'apprentissage des paramètres (1). Cet algorithme simple et universel (2), utilisant le calcul des dérivées partielles a, de plus, permis d'envisager l'utilisation de fonctions de base quelconques vérifiant le critère de dérivabilité.

Dans le domaine de la reconnaissance de parole, l'approche connexionniste est complémentaire de l'approche probabiliste des modèles de Markov.

Les paramètres mis en jeu dans un modèle de Markov doivent être interprétables dans un cadre probabiliste. Ces contraintes limitent les possibilités de modèlisation mais rendent cependant ces systèmes facilement maîtrisables. De plus, l'approche probabiliste permet un alignement temporel du signal de parole en maximisant la probabilité d'émission d'un signal par un modèle.

Dans un réseau connexionniste, aucune contrainte n'est imposée aux paramètres. De ce fait, un tel réseau permet une discrimination explicite des classes entre elles par utilisation de liaisons inhibitrices. Cependant, pour ces systèmes, le problème de l'alignement temporel reste mal maîtrisé.

Ici, nous effectuons l'étude dans un cadre de classification segmentale de phonèmes en "post-processing" d'un modèle de Markov permettant de s'affranchir de ce problème d'alignement. De plus, l'approche segmentale présente l'avantage d'autoriser l'intégration de données de divers types incluant les durées qui, comme le montrent les expériences décrites dans cet article, ont un fort pouvoir discriminant.

La fonction sigmoïdale à support non borné est couramment utilisée pour la non-linéarité des cellules connexionnistes. Quelques auteurs (3) cependant, ont proposé l'utilisation de gaussiennes, fonctions à activation locale. Les expériences suivantes sont menées parallèlement pour des réseaux utilisant des cellules avec l'une ou l'autre des non-linéarités, expériences dans lesquelles les RBF seront progressivement améliorées.

Cet article est organisé de la manière suivante: la section 2 est consacrée à la description des deux types de cellules, et la section 3 présente les techniques d'initialisation et d'apprentissage des MLP et RBF. Dans la section 4, nous décrivons la base de données utilisée. Dans la section 5, nous présentons les premières expériences utilisant l'information segmentale disponible en sortie d'un modèle de Markov, les marqueurs par états et les durées. La section 6 présente l'amélioration des performances apportée par l'utilisation d'un nouvelle paramétrisation pour les données d'entrée. Enfin, la section 7 donne les performances finales obtenues avec les deux types de réseaux.

2. DESCRIPTION DES ARCHITECTURES

Réseaux à unités sigmoïdales.

La sortie d'une cellule sigmoïdale est la somme pondérée des entrées passée à travers une fonction sigmoïde. Son fonctionnement est régi par les équations suivantes:

$$a_i = \sum_{j=0}^{Nj} w_{ij} x_j \qquad x_i = \frac{1}{1+\exp(-a_i)}$$

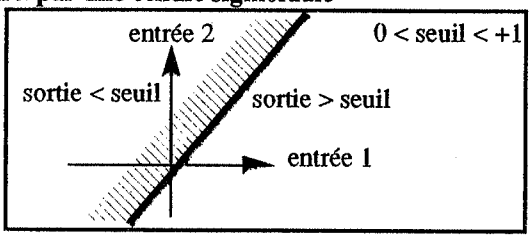
avec x_i sortie de la ième cellule de la couche K (à Ni composantes)

x_j jème composante du vecteur de sortie de la couche K-1 (à Nj composantes)

w_{ij} poids de la connexion reliant la jème cellule de la couche K-1 à la ième cellule de la couche K

Ce type de cellule divise l'espace d'entrée par un hyperplan dont les paramètres sont les poids de la cellule (figure 1).

FIGURE 1. Partition d'un espace d'entrée de dimension 2 par une cellule sigmoïdale



L'architecture utilisée est le MLP (Multilayer Perceptron) à une couche cachée de Nc cellules.

Réseaux à unités gaussiennes.

Le fonctionnement d'une cellule gaussienne à matrice de covariance diagonale est régi par l'équation suivante:

$$x_i = \exp\left(-\sum_{j=0}^{Ne} \frac{(x_j - m_{ij})^2}{(\sigma_{ij})^2}\right)$$

avec x_i sortie

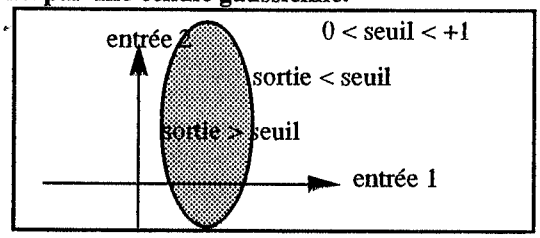
x_j jème composante du vecteur d'entrée

σ_{ij} écart type associé à la jème composante du vecteur d'entrée (à *Ne* composantes)

m_{ij} moyenne associée à la jème composante du vecteur d'entrée

La zone d'activation d'une cellule gaussienne est une ellipsoïde dont le centre et l'étendue sont donnés par les paramètres de la gaussienne. Une telle cellule réagit donc localement sur l'espace d'entrée (figure 2).

FIGURE 2. Partition d'un espace d'entrée de dimension 2 par une cellule gaussienne.



Les réseaux utilisant ce type de cellule sont des réseaux à une couche cachée de Nc cellules gaussiennes et à une couche de sortie de cellules linéaires.

3. INITIALISATION ET APPRENTISSAGE

3.1. Perceptron Multicouche

Les réseaux à cellules sigmoïdales utilisés sont des MLP à une couche cachée. Les poids initiaux sont choisis aléatoirement de manière à ce que les excitations des cellules tombent dans la partie linéaire de la sigmoïde. Les cibles choisies sont du type "1 parmi n" (1 pour la sortie correspondant à la classe du vecteur d'entrée, 0 sinon) et l'apprentissage des poids est effectué par rétropropagation du gradient de l'erreur de Mc Clelland (4) mesurée en sortie. Les pas de gradient sont choisis égaux à 1/*fan-in*, le *fan-in* étant le nombre d'entrées de la cellule considérée.

3.2. Réseau à unités gaussiennes

Les RBF utilisés sont des réseaux à une couche cachée de cellules gaussiennes. L'initialisation du réseau se fait en deux temps:

- *Centres et variances des gaussiennes*: nous appliquons l'algorithme d'affectation d'objet à un groupe de Forgy (5) sur les vecteurs de l'ensemble d'apprentissage. L'initialisation dite "par classe" est utilisée ici: l'algorithme de Forgy est appliqué à chaque classe de vecteurs indépendamment des autres (4 groupes par classe dans les expériences suivantes). La moyenne et l'écart-type d'une gaussienne sont alors initialisés respectivement à la moyenne et à 3 fois l'écart-type des vecteurs tombant dans un groupe issu de la partition de chaque classe.

- *Etage de recombinaison des gaussiennes*: nous calculons la pseudoinverse de la matrice des poids de la couche de sortie de manière itérative. Cette procédure (6) adapte cette matrice à chaque présentation d'un couple entrée-cible de manière à minimiser l'erreur quadratique moyenne en sortie. Dans le cas présent, les vecteurs d'entrée successifs sont les vecteurs moyennes des gaussiennes de la couche cachée. Pour chaque vecteur d'entrée, les cibles dépendent de la classe d'affectation de la gaussienne dont la moyenne est mise en entrée. Deux types de cibles sont proposés: des cibles de type "1 parmi n" égales à 1 pour la classe d'affectation de la gaussienne dont la moyenne est en vecteur d'entrée, 0 sinon, et des cibles tenant compte du nombre de vecteurs de chaque classe rendant maximum chaque gaussienne. Dans ce dernier cas, la

procédure comprend deux étapes:

1. Estimation sur l'ensemble d'apprentissage de *la probabilité conditionnelle qu'un vecteur d'entrée appartienne à la classe C lorsque la gaussienne G est maximale* $P(C,G) = P(x \in C/G\text{maximale})$:

$$P(C,G) = \frac{\sum_{x \in C} 1_{(Gmax)}}{\sum_{x} 1_{(Gmax)}}$$

où la fonction $1_{(Gmax)}$ vaut 1 si G est maximale, 0 sinon.

2. Pour chaque gaussienne G: application en entrée du réseau du vecteur moyenne de la gaussienne G et itération de la procédure de pseudoinversion avec un vecteur cible dont la composante pour la classe C est égale à $P(C,G)$.

Nous pouvons remarquer ici que pour chaque itération dans la procédure de pseudoinversion, la somme des cibles est égale à 1. Les performances des deux types d'initialisation avec en entrée les données du 5.1 sont données dans le tableau 1.

Tab. 1. **Tiers égaux. initialisation RBF**[1]

	N	Nc	Nb ités	Appretissage	Test
RBF 1 parmi N	10	12	0	3.0 (0.20)	**5.1 (0.41)**
RBF "proba"	10	12	0	2.7 (0.20)	**4.4 (0.27)**

La seconde méthode d'initialisation, qui donne de meilleurs résultats, est conservée dans toute la suite de l'étude.

L'apprentissage est effectué par rétropropagation du gradient. Les pas des différentes couches sont choisis pour répondre à un compromis entre adaptation de l'étage de recombinaison des gaussiennes et positionnement des centres de celles-ci.

4. BASE DE DONNEES

La base de données utilisée est un-sous ensemble de la base multilocuteur (environ 800 locuteurs) des chiffres (0 à 9) prononcés isolément à travers le téléphone. Ce sous-ensemble contient les enregistrements des chiffres "5", "6" et "7", segmentés en phonèmes par des modèles de Markov à trois états par phonèmes en alignement forcé. De ces enregistrements sont extraits les parties stables correspondant aux voyelles centrales "in", "i", et "ai". Ceci représente 2114 enregistrements répartis en 720 "in", 689 "i" et 705 "ai".

Les paramétres decrivant le signal sont les 8 premiers coefficients cepstraux plus un coefficient d'énergie et sa dérivée. Ces coefficients sont normalisés entre -1 et +1.

Le choix de ces phonèmes est justifié par le fait que ces trois chiffres se distinguent principalement par leur voyelle centrale et peuvent être confondus en cas de mauvaise discrimination. Ce genre de confusion est fréquemment observé dans les systèmes de reconnaissance.

5. UTILISATION DE L'INFORMATION SEGMENTALE

Une approche segmentale pour la classification de phonèmes permet l'intégration du paramètre "temps". Cette information segmentale est fournie par un modèle de Markov "classique" traitant des trames de 16 ms. Dans les expériences suivantes, nous verrons l'intérêt de ce paramètre pour la discrimination.

MLP et RBF ont un vecteur d'entrée de taille fixe. Les phonèmes à classer étant de longueurs variables, une réduction de données rendant minimale la perte d'information doit être effectuée. Nous testons dans la suite deux segmentations des phonèmes conduisant à cette réduction.

5.1. Segmentation du phonème en tiers égaux

Sans plus d'information sur la réalisation du phonème, la première segmentation utilisée est une segmentation en trois tiers égaux. Le vecteur de dimension 30 résultant de ceci est la concaténation des trois vecteurs qui sont les moyennes calculées sur les trames de chacun des trois segments.

a) Perceptron Multicouche

Les performances d'un tel réseau à 15 cellules cachées sont données dans le tableau 2.

b) Réseau à unités gaussiennes

Les performances des RBF après apprentissage sont données dans le tableau 2.

Tab. 2. **Segmentation en tiers égaux**[1]

	N	Nc	Nb ités	Apprentissage	TEST
MLP	8	15	40	2.6 (0.13)	**4.3 (0.11)**
RBF	10	12	15	1.7 (0.21)	**3.5 (0.21)**

5.2. Segmentation issue des modèles de Markov

Le système de classification étudié est destiné à être utilisé en "post-processing" d'un modèle de Markov. Les informations disponibles contiennent donc les marqueurs de début et fin de phonème ainsi que les marqueurs par états. Supposant que les différents états d'un modèle de Markov mettent en évidence une certaine stabilité du si-

1. Les taux d'erreur sont donnés sous forme de moyenne et écart-type (indiqué entre parenthèses) sur un nombre d'essais N indépendants (initialisation aléatoires différentes) et pour un réseau à Nc cellules cachées. Un écart-type non indiqué signifie la constance des résultats. Une itération équivaut à un passage de l'ensemble d'apprentissage.

gnal, il est intéressant de tester les RBF avec en entrée des vecteurs tenant compte de ces marqueurs d'états. Dans ce paragraphe, le vecteur d'entrée est composé de trois sous vecteurs, chacun d'entre eux représentant *la moyenne des trames tombant dans l'état correspondant du modèle de Markov.*

a) Perceptron Multicouche
Avec cette nouvelle segmentation des phonèmes, les performances d'un MLP sont données dans le tableau 3. La constance des résultats montre la grande robustesse des MLP vis à vis de leur initialisation.

b) Réseau à unités gaussiennes
Les résultats obtenus sont donnés dans le tableau 3.

Ces expériences permettent de remarquer dans un premier temps la bonne stabilité et les bonnes performances des MLP. Ceux-ci sont cependant plus long à converger (de l'ordre de 2 à 10 fois moins rapides en nombre d'itérations. De plus, cette segmentation des phonèmes, amenant de meilleurs performances de classification, permet de penser que les états des modèles de Markov se calent sur des parties stables du phonème, minimisant la perte d'information lors du moyennage. Cette segmentation est conservée pour la suite.

Tab. 3. **Segmentation markovienne**[1]

	N	Nc	Nb ités	Apprentissage	Test
MLP	5	16	60	0.7 (-)	**1.6 (-)**
RBF	10	12	15	1.1 (0.04)	**1.9 (0.14)**

5.3. Intégration de nouvelles données: les durées

L'approche segmentale pour la classification permet l'intégration du paramètre temps en entrée du système, paramètre non accessible dans une approche par trames.
Les paramètres rajoutés aux vecteurs d'entrée sont *les temps passés dans chaque état et la durée totale du phonème*:

- Le temps T passé dans un état est normalisé T^{norm} entre *-0.1* et *+0.1* (ces valeurs ont été expérimentalement choisies (cf 7.1)). Cette normalisation est effectuée en fonction des temps maximum et minimum restés dans un état pour le phonème considéré (T_{max} *et* T_{min}), par la transformation affine donnée par:

$$T^{norm} = 0.2 \cdot [(T-T_{min})/(T_{max}-T_{min})] - 0.1$$

- La durée D du phonème est normalisée par rapport à l'ensemble des phonèmes d'apprentissage. Le coefficient normalisé D^{norm}, compris entre -0.1 et +0.1, est fonction des durées maximale (D_{max}) et minimale (D_{min}) d'un phonème de l'ensemble d'apprentissage. D^{norm} est donné par la transformation affine:

$$D^{norm} = 0.2 \cdot [(D-D_{min})/(D_{max}-D_{min})]-0.1$$

a) Perceptron Multicouche
Les performances de tels réseaux à 20 cellules cachées sur ces nouveaux vecteurs sont données dans le tableau 4.

b) Réseau à unités gaussiennes
Les premiers essais effectués avec un réseau initialisé comme précédemment n'ont pas donné de résultats satisfaisants. Une observation des activations cachées pour les exemples mal reconnus a montré que celles-ci sont toutes très faibles (*<0.05*), soit, dans une partie des gaussiennes où la mise à jour des paramètres ne se fait pas ou peu. Une solution proposée pour résoudre ce problème consiste simplement à initialiser les gaussiennes avec *un écart-type plus grand.* Les résultats donnés dans le tableau 4 ont été trouvés pour un écart-type initial des gaussiennes égal à 4 fois l'écart-type calculé sur les vecteurs d'un même groupe. Cette initialisation donne de meilleurs résultats et est conservée jusqu'à toute nouvelle modification.

Tab. 4. **Durées**[1]

	N	Nc	Nb ités	Apprentissage	Test
MLP	5	20	40	0.7 (0.15)	**0.8 (0.07)**
RBF	10	12	20	0.4 (0.14)	**0.8 (0.14)**

6. MOYENNE CENTRALE, DERIVEE ET DERIVEE SECONDE

Les gaussiennes utilisées dans les expériences sont à matrices de covariance diagonales, pour limiter le nombre de paramètres libres et les temps de calcul. Cependant, ce choix n'est théoriquement consistant que si les coefficients d'entrée sont non corrélés. Précédemment, le vecteur d'entrée était constitué de la juxtaposition de trois vecteurs, chacun de ceux-ci étant la moyenne des trames tombant dans un même état du modèle de Markov. Les coefficients des trames étant corrélés, ceux des vecteurs résultants et ceux du vecteur d'entrée aussi. Une manière simple de résoudre le problème sans toutefois perdre de l'information est de *composer un nouveau vecteur qui contiendra le vecteur moyenne central et les approximations de la dérivée et de la dérivée seconde.*
Si *V1, V2, V3* désignent respectivement les moyennes des trames tombant dans le premier, second et troisième état, l'approximation de la dérivée est choisie

$$D' = (V3 - V1)/2,$$

et l'approximation de la dérivée seconde est choisie

$$D'' = [V2 - (V1 + V3)/2]/2.$$

Les normalisations de ces vecteurs garantissent leurs intervalles de variation bornés entre -1 et +1.
Le nouveau vecteur d'entrée du système sera alors constitué ($V2,D',D''$).

a) Perceptron Multicouche
Des études précédentes ont montré que les réseaux à uni-

tés sigmoïdales offrent de moins bons résultats pour des paramètres décorrélés en entrée. Pour ce type de données, les plans séparateurs interclasses sont orthogonaux aux axes des paramètres d'entrée, plans séparateurs qu'un réseau à unités sigmoïdales a du mal à définir. Cette constatation se trouve confirmée dans les résultats donnés dans le tableau 5 (à comparer avec ceux du tableau 3).

b) Réseau à unités gaussiennes

Nous sommes amenés à reconsidérer le choix de l'écart-type initial des gaussiennes

1. Résultats préliminaires

Les résultats du tableau 5 (RBF base) sont les résultats obtenus avec un RBF incluant les améliorations précédentes. Ces résultats, comparés à ceux du tableau 3, montrent bien l'intérêt de choisir des *coefficients d'entrée le plus décorrélés possibles*, ce qui correspond mieux à des cellules gaussiennes à matrices de covariance diagonales.

2. Choix de la variance initiale des gaussiennes

La procédure de "clustering" utilisée pour l'initialisation est la procédure de Forgy. La distance mesurée pour déterminer le groupe d'appartenance d'un vecteur est la norme de la distance euclidienne entre celui-ci et le centre des groupes. Le fait d'initialiser les gaussiennes avec un écart-type différent de 1 bouleverse la distribution des vecteurs dans les groupes, distribution optimisée par l'algorithme de "clustering". Ceci se traduit par des chevauchements parfois importants de gaussiennes. Cet effet non désirable peut être pallié en prenant la *variance initiale des gaussiennes égale à 1*. La gaussienne associée à un groupe est alors rendue maximale par l'ensemble des vecteurs appartenant à ce groupe. Les résultats présentés (tableau 5) montrent l'amélioration des performances de classification ainsi que de la généralisation des caractéristiques apprises. De plus, nous avons remarqué qu'avec ce choix, le nombre de passages de l'ensemble d'apprentissage nécessaire à la convergence est plus faible.

Tab. 5. **Nouvelles données**[1]

	N	Nc	Nb ités	Apprentissage	Test
MLP	10	20	50	1.3 (-)	**2.1 (-)**
RBF base	10	12	15	0.7 (0.19)	1.1 (0.12)
RBF	10	12	15	0.8 (0.12)	**1.0 (0.14)**

L'amélioration des résultats justifie le choix de variances initiales égales à 1, choix conservé pour la suite.

7. RESULTATS FINAUX

Les performances des MLP étant moins bonnes pour la nouvelle paramétrisation des données d'entrée, nous retiendrons ici leurs meilleures performances obtenues avec les vecteurs moyennes sur les états markoviens plus les informations de durées (tableau 4). Dans la suite nous ne nous interessons plus qu'aux RBF dont les performances peuvent être améliorées. Les dernières expériences sont effectuées sur un vecteur d'entrée comprenant centre, dérivée, dérivée seconde et durées.

1. Essais préliminaires

Les premières expériences pour un RBF de base incluant les choix précédents ont donné les résultats notés dans le tableau 6. Les expériences ont été réalisées aussi pour des durées normalisées entre *-0.5* et *+0.5* (au lieu de *-0.1* et *+0.1*). Dans ce cas, les résultats sont moins bons que précédemment. L'explication de cette sensibilité aux valeurs extrêmes des coefficients est la suivante: lors de la phase initiale d'affectation des vecteurs aux groupes, la distance euclidienne favorise une classification des données en fonction, principalement, de l'axe dominant. Autrement dit, l'influence d'un paramètre sur la partition finale est d'autant plus forte que celui-ci a de fortes valeurs par rapport aux autres. Selon le paramètre concerné, les comportements résultants peuvent être très différents. Le paragraphe suivant donne une solution permettant de s'affranchir de ce problème.

2. Normalisation des données d'entrée

L'hypothèse de départ est que tous les coefficients d'entrée apportent une quantité d'information équivalente. De manière à n'en privilégier aucun lors de la phase de "clustering", *chaque coefficient est normalisé par sa moyenne et K fois son écart-type sur l'ensemble d'apprentissage*. Les coefficients résultants sont tous centrés et d'écart-type *1/K*. Le coefficient *K* est choisi de sorte que la moyenne des activations d'une gaussienne rendue maximale par un vecteur n'appartenant pas à la classe d'affectation de cette gaussienne soit de l'ordre de *0.5* (cette valeur intermédiaire permet un bon comportement en apprentissage). Dans le cas qui nous intéresse, *K=5* (valeur expérimentale). Pour valider l'utilisation de la normalisation des données, les résultats ci-dessous montrent les performances à l'initialisation de deux réseaux, avec et sans cette normalisation:

N=10, à l'initialisation

Sans Normalisation	Avec Normalisation
Test 2.4 (0.58)	Test 0.9 (0.17)

Ces résultats montrent l'efficacité de ce traitement des données d'entrée. La meilleure initialisation obtenue permet une convergence des réseaux très rapide (de 3 à 10 passage de l'ensemble d'apprentissage) pour donner les résultats indiqués tableau 6.

Tab. 6. **Réseau final**[1]

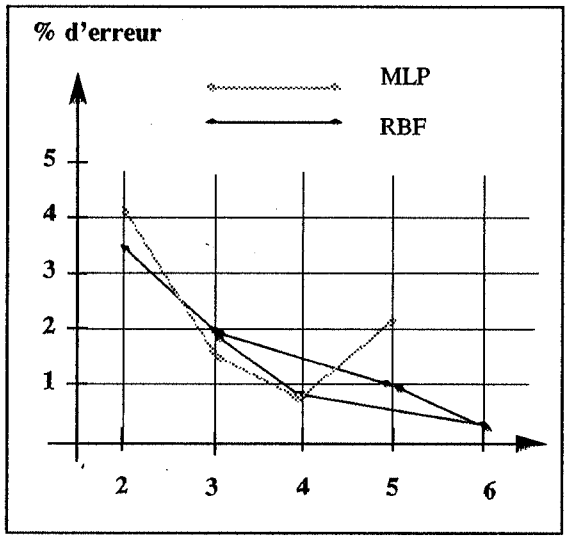
	N	Nc	Nb ités	Apprentissage	Test
RBF base	10	12	15	0.6 (0.12)	**0.5 (0.08)**
RBF	10	12	15	0.4 (0.11)	**0.4 (0.06)**

8. CONCLUSION

1 - Nous avons vu que l'approche segmentale en "post-processing" d'un système de reconnaissance par trames, permet l'intégration du paramètre temps dans les données d'entrée. Les expériences ont montré que ce paramètre améliore la discrimination des phonèmes.
2 - Les différentes expériences menées ont permis de dégager certains problèmes concernant l'utilisation des RBF. Une bonne initialisation de l'étage supérieur et des variances initiales des gaussiennes ainsi que la normalisation des données d'entrée ont rendu le système global à la fois performant en terme de vitesse d'apprentissage (nombre de passages de l'ensemble d'apprentissage) et en terme de taux de reconnaissance.
3 - Le taux de reconnaissance obtenu avec des RBF correctement optimisés sur la tâche est meilleur que celui des MLP. De plus, nous confirmons ici les meilleures performances des RBF à matrices de covariance diagonales pour des données peu ou pas corrélées en entrée, au contraire des MLP qui offrent de meilleures performances avec les données d'entrée corrélées.

La figure 3 fait le bilan des évolutions successives des performances en fonction des expériences et améliorations.

FIGURE 3. Evolution des performances
(l'axe des abcisses suit les tableaux de résultats)



BIBLIOGRAPHIE

1: Y. Le Cun
"Modèles connexionnistes de l'apprentissage", Thèse de doctorat, 1987, Université Paris 6

2: L. Bottou
"Une approche théorique de l'apprentissage connexionniste. Application à la reconnaissance de la parole", Thèse de doctorat, 1991, Paris sud,Centre d'Orsay

3: S. Renals & R. Rohwer
"Learning phoneme recognition using neural networks", Proceedings of IEEE ICASSP 89, Glasgow

4: P. Haffner, A. Waibel, H. Sawai & K. Shikano
"Fast Back-Propagation learning methods for neural networks in speech", Rapport ATR TR-I-0058, 1988.11

5: J.L. Chandon & S. Pinson
"Analyse typologique", Masson édition, 1981

6: T. Kohonen
"Self-organization and associative memory", Springer-Verlag edition, 1983

19es J.E.P. - Bruxelles, 19 au 22 mai 1992.

LE REJET DES ENTREES INCORRECTES D'UN SYSTEME DE RECONNAISSANCE DE LA PAROLE (RAP)

Laurent MAUUARY

CNET/LAA/TSS/RCP, B.P. 40, 22301 Lannion, France
& ACSYS, 6 rue des Coutures, 77200 Torcy, France

Résumé

Ce papier décrit différentes techniques pour rejeter les entrées incorrectes d'un système de reconnaissance de la parole. Une approche consiste à utiliser des modèles "poubelles". Un modèle poubelle est un modèle de Markov Caché (HMM) appris sur un corpus d'énoncés incorrects. La fonction des modèles poubelles est de modéliser, afin de pouvoir le rejeter, ce qui est étranger au vocabulaire. Une autre approche, qui peut être utilisée en association avec les modèles poubelles, consiste à effectuer différentes mesures sur le mot reconnu. A partir de l'information apportée par ces mesures, il faut prendre une décision de rejet ou d'acceptation. C'est un problème de reconnaissance des formes (RF) statistique. Différentes techniques de RF ont été utilisées. Des tests ont été effectués sur des bases de données de laboratoire et aussi sur des bases de données d'exploitation qui fournissent des énoncés incorrects réalistes.

1. INTRODUCTION

Ces toutes dernières années, le domaine de la reconnaissance de la parole a progressé. Des systèmes complets intégrant de la reconnaissance de mots isolés, principalement des serveurs vocaux interactifs (SVI), ont en effet commencé à sortir des laboratoires pour être proposés au grand public, à travers le réseau téléphonique. Ces systèmes sont d'une ambition restreinte, à la mesure du vocabulaire limité qu'ils se donnent pour mission de reconnaître. Mais la mise en exploitation de ces systèmes pose un problème nouveau : **le rejet des entrées incorrectes** (i.e. ne correspondant pas au vocabulaire du système).

Soit parce qu'une erreur de reconnaissance est très coûteuse dans certaines applications, soit parce que le naturel du dialogue favorise des énoncés libres et donc fréquemment hors-vocabulaire, soit parce qu'un environnement bruyant fait que certaines détections ne sont que du bruit, il est crucial de doter ces systèmes d'une procédure leur permettant de refuser de reconnaître un énoncé. Cela donne la possibilité soit d'ignorer l'énoncé, soit d'enclencher une autre procédure demandant la répétition de l'énoncé ou sa confirmation, ou bien posant la question différemment, ou tout simplement mettant un terme au dialogue.

Dans la première partie de cet article, nous présentons le problème du rejet et les techniques utilisées pour prendre la décision rejet/acceptation.

Dans la deuxième partie, consacrée aux expériences de rejet, nous présentons les bases de données utilisées, les mesures de performances du rejet d'un système de reconnaissance et les résultats des expériences de rejet.

2. LE REJET

2.1. EMERGENCE DU PROBLEME

Compte tenu des dégradations apportées par les lignes téléphoniques (limitation de la bande passante, distorsions, ...) et de l'obligation qu'ont les SVI d'être indépendants au locuteur, ils ne peuvent reconnaître qu'un vocabulaire de taille limitée et ils fonctionnent en mots isolés ou connectés. L'analyse des données d'exploitation nous montre que cette limitation à un vocabulaire fermé et précis n'est pas toujours comprise ni respectée par les utilisateurs.

Spitz [1], dans une étude sur l'analyse de données d'exploitation de serveurs téléphoniques à reconnaissance vocale, montre qu'il y a toujours une proportion irréductible d'énoncés incorrects, dépendante de l'application, de la taille du vocabulaire à reconnaître et de la façon dont est mené le dialogue.

Une analyse des appels à Mairievox [2], un serveur téléphonique interactif à commande vocale qui fût en service à Lannion de 1988 à 1991, révèle un fort pourcentage (51 %) de données incorrectes. Dans la même

publication, se trouve également une analyse des appels à Horoscope, un serveur devant reconnaître les douze signes du zodiaque et dont 44 % des entrées sont incorrectes.

Ce qu'il est souhaitable de rejeter est fonction de ce que le système est capable de reconnaître. Dans l'état actuel de la reconnaissance de mots isolés, en mode indépendant du locuteur, nous considérons non-valide tout énoncé contenant plusieurs mots ou bien un seul mot ne correspondant pas au vocabulaire ou bien un seul mot tronqué au début ou à la fin. Nous en arrivons à établir la terminologie suivante pour qualifier un énoncé ou un enregistrement :

valide : contient exclusivement un mot permis
mauvais: contient de la parole ne correspondant pas à un mot permis
bruit : contient autre chose que de la parole
non-valide : mauvais ou bruit

2.2. LA DECISION REJET/ACCEPTATION

Pour rejeter les entrées incorrectes, une approche consiste à utiliser des modèles "poubelles". Un modèle poubelle est un modèle de Markov Caché (HMM) appris sur un corpus d'énoncés incorrects. La fonction des modèles poubelles est de modéliser, afin de pouvoir le rejeter, ce qui est étranger au vocabulaire.

Une autre approche, qui peut être utilisée en association avec les modèles poubelles, consiste à effectuer différentes mesures sur le mot reconnu (et aussi sur le deuxième meilleur mot pour certaines mesures). Imposer un seuil sur une fonction de ces mesures permet de classer les énoncés en valides (f(mesures) > seuil) et en non-valides (f(mesures) < seuil).

2.2.1. LES MESURES UTILISEES

Pour discriminer les énoncés valides des énoncés non-valides, plusieurs mesures ont été utilisées, seules ou en combinaison.

Les deux premières d'entre elles sont le score Viterbi du HMM gagnant (le mot reconnu est celui qui a le meilleur score) et le score sans silence du HMM gagnant. Etant donné que l'approche du CNET [3] est d'utiliser des modèles de silence avant et après les modèles des mots du vocabulaire afin d'éviter une détection explicite des limites précises des mots, le score sans silence est le score du décodage sans prendre en compte les scores relatifs aux modèles de silence. Nous avons utilisé aussi le delta-score, qui est la différence de score entre le score du HMM gagnant et le deuxième meilleur score.

Enfin, pour des modèles par allophones[1], nous avons exploité l'information fournie par la trace du HMM gagnant. La trace est l'ensemble des informations qu'il est possible de recueillir le long du chemin qui représente l'alignement optimal de l'entrée à reconnaître avec le modèle [4]. Dans notre étude, nous avons choisi comme trace le temps passé dans les unités (les phonèmes) d'un modèle par allophones car nous avons pensé que la durée passée dans chaque phonème constituait a priori l'information la plus pertinente. Cette trace est une information à ajouter aux différents scores déjà évoqués (le score du HMM gagnant, le score sans silence du HMM gagnant et le delta-score, disponibles pour toutes les modélisations), afin de prendre une décision rejet/acceptation.

2.2.2. UN PROBLEME DE RECONNAISSANCE DES FORMES STATISTIQUE

Les mesures effectuées sur un mot constituent un vecteur, représentant du mot. Chaque vecteur du corpus d'apprentissage est alors un élément de l'espace des représentations, représentant soit de la classe acceptation (énoncés valides) soit de la classe rejet (énoncés non-valides).

Pour chaque vecteur issu du corpus de test, il s'agit de décider de son appartenance à la classe acceptation ou à la classe rejet. C'est un problème typique de reconnaissance statistique des formes.

2.3. RECONNAISSANCE DES FORMES STATISTIQUE

2.3.1. INTRODUCTION

L'approche statistique en reconnaissance des formes consiste à effectuer sur l'objet à identifier un certain nombre I de mesures, à représenter l'ensemble des mesures par un vecteur de l'espace R^I, et à utiliser des méthodes d'apprentissage statistique pour obtenir une partition de l'espace en classes. Dans la phase de reconnaissance, la classe choisie est celle qui contient le vecteur représentant l'objet à identifier.

Les méthodes de reconnaissance statistique des formes se distinguent selon deux critères [5]. Le premier critère distingue les méthodes bayésiennes (estimation des densités de probabilité pour en déduire selon la théorie bayésienne les surfaces séparatrices) et les méthodes non bayésiennes (construction directe des surfaces séparatrices). Le deuxième critère distingue entre les méthodes paramétriques (recherche de la meilleure densité, ou surface, à l'intérieur d'une famille de densités ou de surfaces) et les méthodes non paramétriques (pas d'hypothèses a priori sur les lois des classes ou les surfaces séparatrices).

[1] Il est bien entendu possible d'utiliser une trace pour d'autres types de modélisations, par exemple le temps passé dans chaque état pour un modèle par mots, mais nous n'avons utilisé une trace que pour les modélisations par allophones qui donnent les meilleurs taux de reconnaissance.

2.3.2. UNE METHODE PARAMETRIQUE BAYESIENNE

Pour estimer les lois de probabilité que suivent les classes, l'approche paramétrique bayésienne consiste à choisir a priori un modèle pour chacune des densités de probabilité des classes et à se contenter d'estimer les paramètres de ce modèle.

Dans cette étude, l'hypothèse toujours retenue a été la normalité des lois. Si cette hypothèse simplificatrice est parfois très grossière, la simplification est très importante : il va suffire d'estimer deux paramètres qui sont un vecteur moyenne et une matrice de covariance.

Dans le problème du rejet, comme il n'y a que les deux classes rejet et acceptation, cette approche a toujours conduit à utiliser comme fonction sur les mesures le logarithme du rapport de vraisemblance. Nous définissons le rapport de vraisemblance comme étant le rapport des fonctions de densités de probabilité des deux classes.

2.3.3. LES PLUS PROCHES VOISINS

La méthode des k plus proches voisins (ppv) est une méthode non paramétrique non bayésienne qui vise à déterminer directement la partition de l'espace R^I en classes, sans faire d'hypothèses sur la nature des distributions sous-jacentes, ni sur la nature des surfaces séparatrices idéales.

Soit x un point de l'espace R^I que l'on souhaite reconnaître. La règle de décision consiste à décider que x appartient à la classe la plus représentée parmi les k points de l'ensemble d'apprentissage les plus proches de x. Cette méthode nécessite le choix du nombre de ppv et d'un type de distance afin de calculer les ppv.

Pour cette approche, la fonction sur les mesures est donc le nombre de voisins de la classe acceptation.

3. EXPERIENCES DE REJET

3.1. BASES DE DONNEES

Les bases de données, utilisées pour l'apprentissage des modèles et pour les tests de rejet, sont de deux types qualifiés *de laboratoire* et *d'exploitation*.

3.1.1. DONNEES DE LABORATOIRE

Les bases de données de parole dites "de laboratoire" sont constituées d'enregistrements provenant de locuteurs volontaires, prévenus et coopératifs. Les données sont issues de séances d'enregistrement à travers le réseau téléphonique (principalement en interurbain), et non pas d'appels réels à un quelconque serveur vocal. Les mots du vocabulaire de l'application visée sont tous prononcés un nombre égal de fois par chaque locuteur.

Le tableau ci-dessous représente la composition des corpus d'apprentissage et de test pour 3 des bases de données de laboratoire.

	apprentissage	test
Trégor (36 mots)	8354	8400
Chiffres (10 mots)	2065	2080
500 mots	9766	4883

3.1.2. DONNEES D'EXPLOITATION

Les bases de données de parole dites "d'exploitation" sont constituées d'enregistrements provenant d'appels réels d'usagers à un serveur vocal interactif en exploitation. Les mots du vocabulaire de l'application n'y sont donc pas forcément représentés équitablement. De plus, les données ne sont pas forcément des mots du vocabulaire, soit à cause des usagers qui ne respectent pas toujours les consignes, soit à cause de l'environnement sonore qui est rarement idéal.

L'utilisation conjointe des données d'exploitation et de laboratoire trouve sa justification dans [6] où la preuve est faite que l'apprentissage des modèles avec ces deux types de données fournit les meilleures performances pour les SVI. La base de données Horoscope, dont le vocabulaire est constitué des douze signes du Zodiaque, contient ces deux types de données.

Le tableau ci-dessous indique la répartition des enregistrements valides et non-valides dans les données de laboratoire et d'exploitation et dans les corpus d'apprentissage et de test. Les entrées non-valides sont constituées principalement de bruits et de mots hors vocabulaire.

	laboratoire	exploitation	apprentissage	test
valides	1567	3604	2664	2507
non-valides	132	1738	869	1001

3.2. MESURES DES PERFORMANCES

Un enregistrement valide peut être :
- correctement reconnu,
- correctement reconnu et rejeté (Faux Rejet),
- mal reconnu (Erreur de Substitution),
- mal reconnu et rejeté (Erreur de Substitution Rejetée).

De même, un enregistrement non-valide peut être :
- rejeté,
- (mal) reconnu (non-rejet, ou Fausse Acceptation).

La performance sur les enregistrements valides est mesurée par la somme du taux d'erreurs de substitution (taux ES), du taux de faux rejet (taux FR) et du taux d'erreurs de substitution rejetées (taux ESR) :

$$ES = \frac{\text{nbre d'erreurs de substitution}}{\text{nbre d'enregistrements valides}}$$

$$FR = \frac{\text{nbre de faux rejets}}{\text{nbre d'enregistrements valides}}$$

$$ESR = \frac{\text{nbre d'erreurs de substitution rejetées}}{\text{nbre d'enregistrements valides}}$$

La performance sur les enregistrements non-valides est mesurée par le taux de fausses acceptations (taux FA):

$$FA = \frac{\text{nbre de fausses acceptations}}{\text{nbre d'enrg. non-valides}}$$

Comme les deux performances résultent d'un compromis, les résultats seront présentés sous forme de courbe avec le taux FA en ordonnée (**% d'erreurs sur les NON valides**), et la somme des taux ES, FR et ESR en abscisse (**% d'erreurs sur les valides**).

3.3. MODELISATION DES TRACES

La trace est une information sur l'alignement du mot avec le modèle. Dans cette étude, elle n'a été utilisée que pour les modélisations par allophones. Pour un mot du vocabulaire modélisé par N phonèmes, la trace est donc un ensemble de N nombres. Pour les N-1 premiers nombres, la valeur pour un phonème est le temps passé dans ce phonème (c'est-à-dire le nombre de trames de l'enregistrement mises en correspondance avec les gaussiennes qui modélisent ce phonème). La valeur pour le dernier nombre est la durée totale passée dans tous les phonèmes du mot.

Pour chaque mot du vocabulaire, une modélisation statistique est effectuée pour les classes acceptation et rejet.

Cette modélisation nécessite un nombre minimum de traces. Or le propre des modèles par allophones est d'autoriser plusieurs réalisations phonémiques différentes pour chaque mot du vocabulaire. Mais certaines réalisations sont peu fréquentes et la trop petite taille des bases de données ne permet pas de les avoir en nombre suffisant pour les modéliser. Il nous a donc fallu observer les différentes réalisations générées sur les bases de données et effectuer un traitement afin d'obtenir un seul type de trace pour chaque mot du vocabulaire. Nous constatons que les variations correspondent toujours à la présence éventuelle d'un e muet après une consonne et avant une pause ou entre deux consonnes. Le traitement consiste donc à ne pas prendre en compte ce phonème quand il est présent.

3.4. RESULTATS

Un logiciel de modélisation statistique [3] a été utilisé pour la création des modèles, l'apprentissage de ces modèles et les tests de reconnaissance.

3.4.1. TESTS SUR LE TREGOR

Le vocabulaire à reconnaître est celui du Trégor; le vocabulaire à rejeter est celui des Chiffres.

Une description détaillée des différentes modélisations utilisées se trouve dans [7]. Précisons simplement que pour les modèles par allophones, 27 coefficients ont été utilisés (8 coefficients cesptraux, plus l'énergie et les dérivées premières et secondes de ces 9 coefficients). Pour les autres modèles, 8 coefficients ont été utilisés (6 coefficients cesptraux, l'énergie et sa dérivée).

La figure 1 est une étude de l'influence du type de modélisation : modèles par mots (*mot*), modèles par pseudo-diphones (*psd*), modèles par mots et poubelles (*mot+poub*), où 3 modèles poubelles ont été appris avec la base de données des 500 mots. *Ssl* (resp. *dsc*) signifie l'utilisation d'un seuil sur la valeur du score sans silence (resp. du delta-score), la fonction sur ces mesures est alors l'identité.

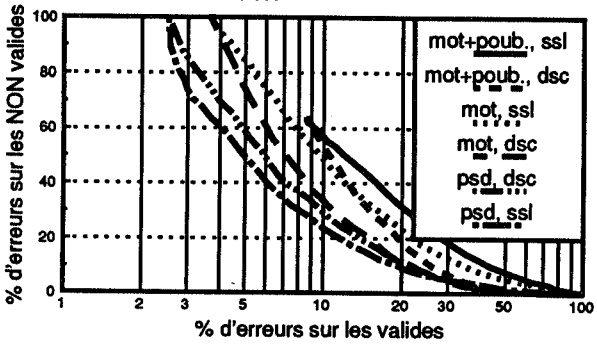


Figure 1 **Etude de différentes modélisations du vocabulaire Trégor**

La première conclusion concerne les modélisations sans poubelles : plus un modèle est performant (taux de bonne reconnaissance) plus il permet de faire du rejet ce qui revient à dire que mieux un vocabulaire est modélisé plus il est facile de détecter des entrées qui ne sont pas des mots du vocabulaire.

La seconde concerne l'apprentissage des modèles poubelles. Dans le cas présent, ils ont été appris avec la base de données des 500 mots. Nous constatons que cet apprentissage n'est pas adapté au rejet des chiffres car la diminution du taux de reconnaissance n'est pas compensée par la modélisation (insuffisamment bonne) du vocabulaire à rejeter.

Enfin, le *ssl* et *dsc* sont des mesures d'une efficacité similaire pour faire du rejet.

Suite à cette première constatation, les tests suivants ont été effectués sur des modèles par allophones qui donnent les meilleurs taux de reconnaissance. Le modèle utilisé correspond à la structure des allophones la plus performante décrite dans [8].

Sur la figure 2, nous retrouvons deux des mesures de la courbe précédente : le *ssl* et le *dsc*. Ces deux mesures peuvent être utilisées en combinaison, le plus simple étant d'utiliser les connecteurs logiques *et* et *ou* : le mot reconnu est rejeté s'il est rejeté par la mesure *ssl et/ou* la mesure *dsc*. La trace a été utilisée avec le logarithme du rapport de vraisemblance (*lrv(trace)*) comme fonction sur la trace. Cette fonction est elle-même une mesure et a été utilisée avec le *ssl*, le *dsc* et le *asl* (score avec silence) avec le *lrv* comme fonction sur ces mesures (*lrv(asl,ssl,dsc,lrv(trace))*). Le *lrv(asl,ssl,dsc,lrv(trace))* est calculé globalement pour

tous les mots du vocabulaire (pour un meilleur respect de l'hypothèse de normalité) et non pas pour chaque mot du vocabulaire comme c'est le cas pour le *lrv(trace)*.

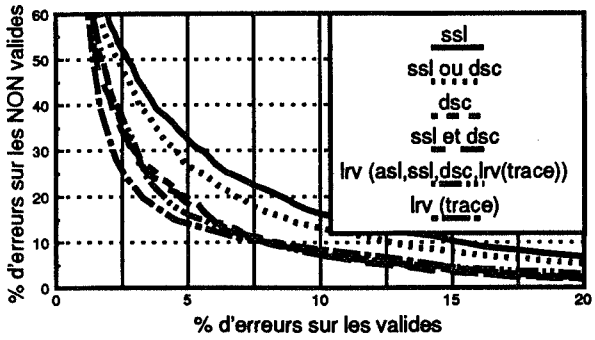


Figure 2 **Comparaison de différents critères de décision sur une modélisation par allophones du vocabulaire Trégor**

La trace est une information pertinente pour faire du rejet, puisqu'on obtient pratiquement les meilleurs résultats avec cette mesure seulement. Si l'utilisation conjointe des 4 mesures *asl*, *ssl*, *dsc* et *lrv(trace)* ne donne pas les meilleurs résultats, bien qu'utilisant le maximum d'informations, c'est que l'hypothèse de normalité n'est pas respectée pour le *dsc* et surtout pour le *lrv(trace)*, ce qui ne permet pas d'exploiter au mieux l'information que constituent ces 4 mesures.

Nous nous sommes ensuite intéressés à l'influence des modèles poubelles (figure 3). Un modèle poubelle a été ajouté au modèle par allophones de la figure 2.

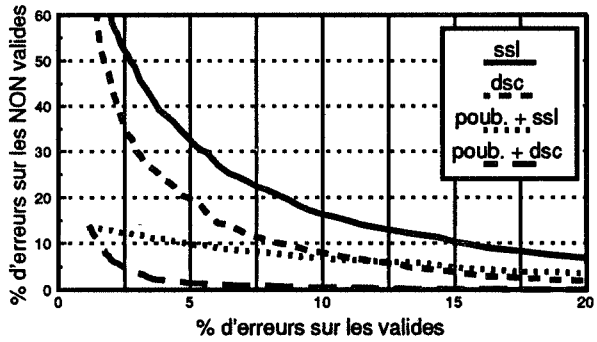


Figure 3 **Influence des modèles poubelles sur une modélisation par allophones du vocabulaire Trégor**

La figure 1 ayant illustré la difficulté de l'apprentissage des modèles poubelles, le modèle poubelle utilisé a été appris avec la base de données des Chiffres pour être sûr de bien modéliser les entrées à rejeter.

Nous constatons la grande efficacité de ce procédé. Comme cela n'apparaît pas sur la figure, nous précisons que le taux d'erreur de reconnaissance est passé de 0,92 à 1,15% en ajoutant le modèle poubelle. Cette baisse minime du taux de reconnaissance est largement compensée par une très bonne modélisation du vocabulaire à rejeter. Cependant la grande difficulté de cette approche est de bien connaître les entrées à rejeter, ce qui n'est pas du tout évident dans la pratique. Comme les bases de données d'exploitation permettent d'obtenir des énoncés non-valides réalistes, la suite des tests a été effectuée sur la base de données Horoscope.

3.4.2. TESTS SUR HOROSCOPE

La figure 4 correspond à l'utilisation de différents critères de décision avec une modélisation par allophones. La nouveauté est l'utilisation des ppv sur les mesures *asl*, *ssl*, *dsc* et *lrv(trace)*. Nous avons considéré les 9 ppv, au sens de la distance euclidienne. Afin de ne favoriser aucun axe pour le calcul de la distance euclidienne, les écart-types de chacun des axes ont été rendus unitaires.

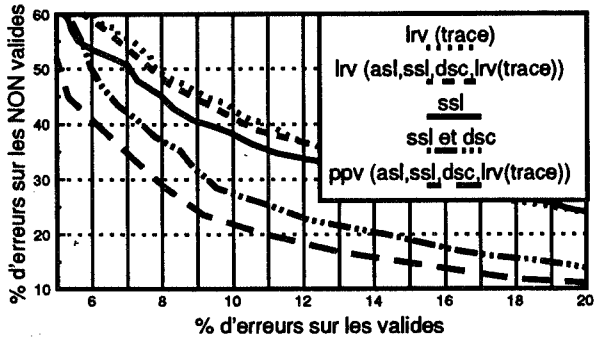


Figure 4 **Comparaison de différents critères de décision sur une modélisation par allophones du vocabulaire Horoscope**

La grande différence avec les tests effectués sur les bases de données de laboratoire est que l'information que représente la trace est beaucoup moins pertinente quand on utilise des énoncés non-valides réalistes.

La méthode des *ppv(asl,ssl,dsc,lrv(trace))* donne de meilleurs résultats car elle ne nécessite pas d'hypothèses et que seuls les ppv du mot reconnu sont considérés, par opposition à l'apprentissage global effectué pour le *lrv(asl,ssl,dsc,lrv(trace))*.

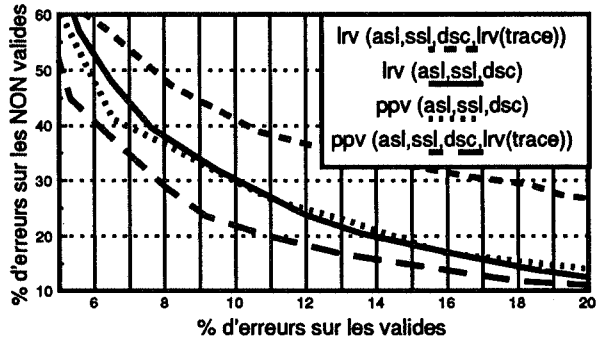


Figure 5 **Utilité de l'information apportée par les traces**

Nous nous sommes ensuite demandé ce que les traces peuvent représenter comme information utile (non redondante par rapport aux mesures *asl*, *ssl* et *dsc*). Les performances (figure 5) des décisions re-

jet/acceptation avec et sans l'utilisation de cette information (*lrv(trace)*) ont donc été comparées.

Les résultats obtenus permettent de conclure à l'utilité de l'information apportée par les traces. En effet, nous constatons une amélioration pour la méthode des ppv. Par contre il y a dégradation pour la méthode lrv avec laquelle nous ne parvenons pas à tirer profit de l'information apportée par le *lrv(trace)*.

Nous nous sommes ensuite intéressés à l'utilisation de modèles poubelles (figure 6). Deux modèles poubelles ont été ajoutés à la modélisation par allophones. Un des modèles a été initialisé avec les non-valides mauvais et l'autre avec les non-valides bruits. L'apprentissage a été laissé libre avec les non-valides mauvais et bruits.

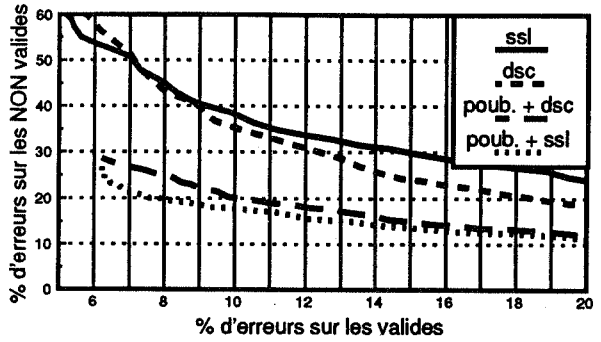


Figure 6 **Influence des modèles poubelles sur une modélisation par allophones du vocabulaire Horoscope**

Cette méthode se révèle très efficace. Le taux d'erreurs de reconnaissance est passé de 3,75 à 6,22% alors que le taux de fausses acceptations est tombé de 100 à 28,67%. Les mesures *ssl* et le *dsc* donnent des résultats très similaires.

Il nous restait donc à tester différents critères de décision sur une modélisation par allophones plus modèles poubelles du vocabulaire Horoscope (figure 7).

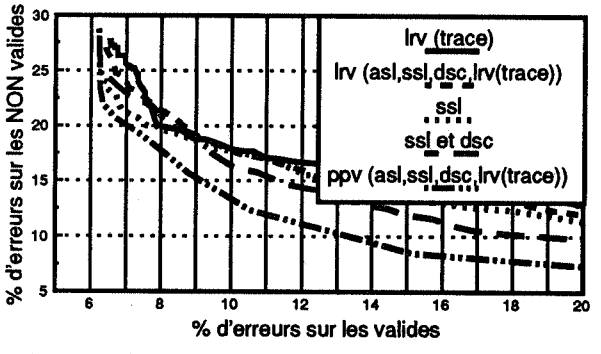


Figure 7 **Comparaison de différents critères de décision sur une modélisation par allophones, avec modèles poubelles, du vocabulaire Horoscope**

Comme pour la modélisation par allophones sans modèles poubelles, le *ppv(asl,ssl,dsc,lrv(trace))* donnent de meilleurs résultats que le *lrv(asl,ssl,dsc,lrv(trace))*. L'utilisation conjointe de modèles poubelles et de la méthode des ppv permettent de rejeter **80 %** des entrées incorrectes avec un taux de reconnaissance de **93 %** sur des bases de **données d'exploitation** et un vocabulaire de 12 mots.

4. CONCLUSION

Les tests effectués sur les bases de données de laboratoire ont mis en évidence la nécessité d'obtenir une bonne modélisation du vocabulaire afin de pouvoir rejeter les entrées incorrectes. Il a été montré aussi que la trace (une information de durée seulement dans cette étude) est une information pertinente pour faire du rejet sur les données de laboratoire et que les modèles poubelles peuvent être efficaces s'ils sont correctement appris. Les tests effectués sur les données d'exploitation ont montré que la trace est alors une information moins pertinente que pour les données de laboratoire mais reste utile, quoique difficile à utiliser car n'ayant donné satisfaction que pour la méthode des ppv, une technique de RF ne faisant pas d'hypothèses a priori mais coûteuse en phase de reconnaissance. Une solution possible consiste à utiliser des classificateurs connexionnistes, qui ne nécessitent pas d'hypothèses à priori, et qui sont très peu coûteux en phase de reconnaissance.

5. REFERENCES

[1] J. Spitz : "*Collection and analysis of data from real users : implications for speech recognition/understanding systems*"; DARPA Workshop 1991, Asilomar, Février 1991.

[2] L. Mathan, D. Morin : "*Speech field databases : development and analysis*"; EUROSPEECH'91, Genova, Septembre 1991.

[3] C. Gagnoulet, D. Jouvet : "*Reconnaissance de la parole et modélisation statistique : expérience du CNET*"; Traitement du Signal, Vol. 7, n° 4, 1990.

[4] L. Mathan : "*Contributions à la reconnaissance de la parole pour les serveurs vocaux interactifs*"; thèse ENST, 1991.

[5] G. Gaillat : "*Reconnaissance des formes : méthodes statistiques*"; 1983 Editions ENSTA.

[6] D. Morin : "*Influence of field data in HMM training for a vocal server*"; EUROSPEECH'91, Genova, Septembre 1991.

[7] D. Jouvet : "*Application des modèles de Markov à la reconnaissance de la parole*"; 2èmes Journées Nationales GRECO-PRC Communication Homme-Machine, Toulouse, 29-30 Janvier 1991.

[8] D. Jouvet, K. Bartova, J. Monné : "*On the modelization of allophones in an HMM based speech recognition system*"; EUROSPEECH'91, Genova, Septembre 1991.

19es J.E.P. - Bruxelles, 19 au 22 mai 1992.

DES LEXIQUES AUX RÈGLES : VERS UNE MÉTHODE DESCRIPTIVE DE LA PHONÉTISATION DU FRANÇAIS

R. BELRHALI L. LIBERT V. AUBERGÉ & L-J. BOË

ICP URA CNRS n° 368 INPG/ENSERG UNIVERSITÉ STENDHAL
BP 25 38 040 GRENOBLE CEDEX 9 FRANCE

Résumé

Notre étude s'inscrit dans le cadre de l'analyse linguistique des phénomènes de phonétisation du français. Une grammaire initiale, écrite en langage TOPH [Aubergé, 1985], a permis de phonétiser *Le 60 000* de l'ICP. La transcription phonétique obtenue de chaque caractère, systématiquement vérifié dans *Le Petit Robert 1* (1990), a permis l'élaboration de nouvelles règles. Cet ensemble confère à la grammaire un taux d'efficacité voisin de 100% sur le dictionnaire.

1. INTRODUCTION

Pourquoi une nouvelle étude sur la phonétisation alors que de nombreux travaux ont été réalisés dans ce domaine ? Ils couvrent déjà : la description linguistique à des buts didactiques [Catach, 1984 ; Véronis, 1987 ; Lacheret-Dujour, 1990], l'étude des variantes phonétiques pour la reconnaissance, correction orthographique [Lahens, 1987] et la bien sûr synthèse à partir du texte [Divay et Guyomard, 1977 ; Prouts, 1980 ; Laporte, 1988].

Dans notre cas il s'est agit de proposer un modèle automatisable qui se prête directement à une lecture linguistique de la phonétisation. Dans ce module, les différentes composantes linguistiques seront organisées, identifiées et répertoriées. Le choix du formalisme TOPH [Aubergé, 1991] permet un compromis entre représentation par règles et représentation par lexiques, tout en imposant une méthodologie d'exploration systématique des correspondances orthographiques-phonétiques de toutes les sous-chaînes, du graphème au mot. L'organisation entre règles et lexiques sera guidée par des critères linguistiques.

Nous présentons ici la méthodologie systématique, appliquée pour cerner exhaustivement les correspondances mises en œuvre dans un dictionnaire représentatif du français, *Le 60 000* de l'ICP. Nous essaierons ensuite de classer ces correspondances selon le niveau linguistique des informations véhiculées à l'oral (graphémique, morpho-phonémique, lexical, catégoriel, syntaxique et sémantique). Les solutions proposées ne nécessitent pas l'explicitation des attributs de chacun de ces niveaux linguistiques. Finalement, nous présenterons les caractéristiques de la grammaire élaborée dans un environnement HyperCard.

2. MÉTHODOLOGIE

Le noyau initial [Aubergé, 1985] est constitué d'une grammaire de règles de base dans lesquelles ont été identifiés les types de lexiques nécessaires. Cette première description a été utilisée pour phonétiser *Le 60 000* (qui contenait à l'origine les mots dans leur seule forme orthographique). Chaque forme phonétique a ensuite été systématiquement vérifiée et corrigée sur la référence du *Petit Robert 1*.

Ce nouveau matériau a été analysé pour étendre la grammaire par l'ajout d'irrégularités et la complétion des lexiques [Belrhali *et al.* 1990]. Il faut souligner que la classification des lexiques à l'intérieur des règles, elles-mêmes structurées, relève de deux démarches *a priori* indépendantes.

• L'organisation logique : chaque fois qu'il est possible, les règles (et lexiques) ont été regroupées et

fusionnées. Nous décrirons au paragraphe suivant ce que signifie, en langage TOPH, cette notion de concision.

• L'organisation linguistique : la structure d'une règle correspond à un mécanisme linguistiquement identifié et les éléments d'un même lexique partagent le même comportement linguistique.

L'un des points essentiel est la superposition de ces deux démarches : l'organisation logique est en général facilement interprétée par des critères linguistiques.

3. LE LANGAGE TOPH

TOPH, langage et environnement logiciel, a été développé à l'ICP pour la synthèse multilingue [Aubergé *et al*, 1987]. Il réalise la transcription de textes orthographiques sans contraintes sur leur forme (Figure 1). Il faut noter, au passage, que dans l'environnement HyperCard choisi pour développer la grammaire du français et la base de données du *60 000*, les codes phonétiques utilisés sont bien ceux de l'API.

Une grammaire en langage TOPH comporte :

(1) une déclaration d'ensembles, éventuellement vide

(2) une partie règles, où peuvent être appelés les ensembles préalablement définis.

Bien que décrits dans une syntaxe identique, les ensembles peuvent avoir des fonctions linguistiques différentes :

- identification typographique ; un ensemble "#" de séparateurs de mots peut être déclaré par :
  "#" = (; , . , ! , ? , /(, /) , - , /)
- classe orthographique, par exemple :
  "Consonnes nasales" = (n, m)
- lexique d'irrégularités :
  "Lexique des "-s" prononcés en finale" = (a , bu)

Une règle décrit l'association d'une chaîne d'entrée quelconque – dont l'unité significative minimale est la sous-chaîne orthographique (graphèmes) – et d'une chaîne de sortie quelconque (sons). Les règles sont regroupées par classe, une classe étant définie par le premier caractère de la chaîne à transcrire. Lors de la transcription, les règles seront examinées selon l'ordre dans lequel elles sont écrites ; la première règle validée est appliquée sur le texte d'entrée. Les contextes gauche et droit de la chaîne à transcrire pondèrent éventuellement la réécriture. Un contexte se définit par :

– une sous-chaîne graphémique *a*
– un ensemble *b*
– l'énumération de *a* ou *b* (notée "*a, b*")
– la concaténation de *a* et *b* (notée "*a* + *b*")

L'exemple qui suit met en évidence la notion de concision abordée au paragraphe précédent : les deux premières règles sont fusionnées en une seule, puis l'énumération dans cette règle est identifiée par un lexique, finalement appelé dans la dernière règle :

("#" + cubitu) +s+ ("#") = [s]
("#" + huméru) +s+ ("#") = [s]
("#" + radiu) +s+ ("#") = [s]
ou ("#" +cubitu,huméru,radiu) +s+ ("#") = [s]
ou ("#" + "LEXIQUE : S. FINALE") +s+ ("#") = [s]
(lexique correspondant au graphème "s" prononcé en position finale).

On s'aperçoit finalement que dans ce type de grammaire terminale, la différence entre règle et lexique est floue puisque la plupart des mécanismes peuvent s'interpréter comme des mises en correspondance (sous contrainte contextuelle) entre forme graphémique et forme phonétique, et donc comme des lexiques de correspondance "au" = [o] aussi bien que "as" = [as].

4. CLASSIFICATION DES MÉCANISMES ET SOLUTIONS

La phonétisation véhicule des informations linguistiques de différents niveaux : graphémique, lexical, morpho-phonémique, catégoriel, syntaxique et sémantique.

4.1. Le niveau graphémique

Il correspond à la réécriture régulière d'une chaîne graphémique en une chaîne phonétique. C'est donc un niveau phonotactique pour lequel les règles ne tiennent pas compte de la forme globale du mot ni *a fortiori* de niveaux plus complexes. Par exemple :

- la chaîne "eau" se transcrit [o]
- si "s" est entre deux voyelles, il se réécrit [z] sinon "s" se réécrit [s]

4.2. Le niveau morphophonémique

Les unités lexicales du français sont composées au moins d'affixes et d'une base lexicale : les règles de frontières de mot s'appliquent en général aux frontières de morphèmes. C'est ainsi qu'un mot graphique devra être considéré, au niveau de la phonétisation, comme autant d'unités morphémiques :

4.2.1. Préfixe + Base

Par exemple :

- "asocial" est prononcé comme "a" + "social" : "s" se transcrit [s]
- "polyacide" est prononcé comme "poly" + "acide" : "ya" se transcrit [ia] au lieu de [ja]

4.2.2. Base + Suffixe

Par exemple le problème de la polyphonie de "tie, tion", où "t", lorsqu'il n'est pas précédé de "s", se transcrit [s] ou [t]. On peut supposer *a priori* que "t" se prononce [s] lorsque "tie" et "tion" sont suffixes. Mais lorsque ces suffixes sont des flexions, on retombe sur le problème des ambiguïtés catégorielles (voir § 4.4.). Par exemple :

- "t" se prononce [s] dans "acrobatie", et [t] dans "sotie"

Lorsque l'on analyse [Aubergé, 1985] des listes tirées du *Dictionnaire Phonétique inverse du français*

de Juilland ou du *60 000*, on s'aperçoit que cette règle n'est pas toujours vérifiée. Pour ce type de problème, nous proposons des solutions systématiques par lexiques.

4.2.3. Base + Base

L'agglutination est une construction marginale en français, mais nous en retrouvons des traces :

- "soubresaut" est prononcé comme "soubre" + "saut" "tournesol" est prononcé comme "tourne" + "sol"

Il en est de même dans les constructions qui contiennent des radicaux grecs fonctionnant comme préfixes :

- "immunosuppresseur" est prononcé comme "immuno" + "suppresseur"
- "chimiosynthèse" est prononcé comme "chimio" + "synthèse"

4.3. Le niveau lexical

Un ensemble de mots peut partager une même particularité de réécriture (mots d'emprunt de même origine, de même usage...). Ces ensembles de mots irréguliers seront réunis dans des lexiques. Il arrive que certains mots n'appartiennent à aucun lexique. C'est pourquoi nous qualifierons ces mots d'exceptions par opposition aux irrégularités. On peut donc établir une liste de tous ces lexiques et vérifier qu'ils sont identifiables à des classes linguistiquement explicables.

Nous présentons au § 5 la liste et la taille des principaux lexiques linguistiques qui ont été établis avec *Le 60 000* de l'ICP. Le contenu de chacun de ces lexiques est donné dans [Belrhali et Libert, 1991]. À quelques rares exceptions, il s'agit de mots d'emprunt ou de mots "savants". Par exemple :

- "ch" se prononce [k] dans les mots savants, en général d'origine grecque, "trachéal, trachéen, trachéite, trachéostomie, trachéotomie, trachome, trachyte"
- "er" se prononce [e] lorsqu'il est une flexion verbale ("chanter"), ainsi qu'en finale des noms et adjectifs polysyllabiques ("boulanger, léger")
- "er" se prononce [ɛʀ] en finale des noms et adjectifs monosyllabiques, comme dans "cher"..., ainsi qu'en finale de mots savants ou d'emprunt, comme dans "africaner"...
- "er" se prononce [œʀ] en finale de mots d'emprunt, comme dans "angledozer, bitter, bookmaker"

Les exceptions sont peu nombreuses et dues à des accidents diachroniques ou des faits d'usage. Par exemple :

- "ch" se prononce [ʃ] dans "trachée", son usage plus courant l'a différencié des mots de même origine où "ch" se prononce [k] ("trachéotomie"...)
- "pt" se prononce couramment [pt] dans "dompter", alors que sa prononciation référencée dans *Le Petit Robert* est [t]. Ce mot reflète un accident diachronique : son orthographe s'est transformée arbitrairement de "domter" (du latin "domitare") à "dompter", par suite d'une analogie au mot "compter" et l'usage n'a pas suivi la norme de prononciation
- *Le Robert oral-écrit* fait état de "dilemme" qui se prononce [dilɛm] mais aussi [dilɛmn] comme "indemne" [ɛ̃dɛmn]
- "carrousel", emprunté au XVII^{e} siècle à l'italien "carosello", a conservé la prononciation de s en [s] et se prononce [kaʀusɛl]

4.4. Le niveau catégoriel

Du point de vue de la phonétisation, quelques unes des marques graphiques fonctionnelles (dérivations et flexions) sont ambiguës au niveau strictement phonotactique. En particulier, les finales "ent", "tions" ont des prononciations différentes selon la catégorie lexicale du mot :

- "ent" ne se prononce pas lorsqu'il est la flexion verbale de la troisième personne du pluriel ("chantent")
- "ent" se prononce [ã] en finale de noms, d'adjectifs, d'adverbes et de quelques formes verbales ("vent, lent, souvent, sent")
- "tions" se prononce [tjɔ̃] lorsqu'il est une flexion verbale ("chantions"), sauf dans "balbutions, argutions, initions"
- "tions" se prononce [sjɔ̃] en finale de noms au pluriel ("rations")

Si une description linguistique consiste à lever ces ambiguïtés par la valeur catégorielle du mot considéré, une solution pratique est de constituer des lexiques de formes « non verbales », lorsqu'elles ne sont pas homographes (cf. § 5). Mais lorsque les deux unités sont homographes, à l'ambiguïté catégorielle s'ajoute la lexicale qui ne peut pas être levée par la grammaire TOPH puisqu'il faut procéder à une analyse syntaxique. Par exemple :

- "er" : "se fier [fje]" (verbe) *vs.* "fier [fjɛʀ]" (adjectif)
- "ent" : "président [pʀezid]" (verbe) *vs.* "président" [pʀezidɑ̃]" (nom)
- "tions" : "portions [pɔʀtjɔ̃]" (verbe) *vs.* "portions" [pɔʀsjɔ̃]" (nom)
- ambiguïté flexionnelle : "un os [œn-ɔs]" (singulier) *vs.* "des os [dez-o]" (pluriel)

4.5. Le niveau sémantique

Il existe enfin un dernier type d'ambiguïté qui ne peut être levé par la syntaxe, car il y a homomorphie lexicale et catégorielle. La décision doit être prise alors au niveau sémantique, voire pragmatique. C'est le cas pour :

- "les fils [fis]" (« enfants ») *vs.* " les fils [fil]" (« à coudre »)
- "jet [jɛ]" (« jet d'eau ») *vs.* " jet [jɛt]" (« avion »)

5. UNE PROPOSITION POUR UN MODÈLE DE PHONÉTISATION DU FRANÇAIS

Dans la grammaire, nous n'avons pas inclus le lexique des mots dont la finale en "-ent" se prononce [ɑ̃], en raison de sa taille trop importante (2 680 mots) par rapport à l'ensemble des autres lexiques. Mis de côté pour l'instant, il sera remplacé par le lexique des formes verbales en "-ent", que nous n'avons pas encore rassemblé (mais qui sera nettement moins volumineux). La grammaire est actuellement constituée d'un ensemble de 1 270 règles dont une vingtaine activent environ vingt lexiques contenant 1 650 mots.

Bien qu'il soit difficile de décider si une règle définit une exception ou une réécriture régulière, on peut estimer que environ 600 d'entre elles décrivent à peu près 900 exceptions et les 650 autres caractérisent le mécanisme de base de la réécriture.

Les lexiques appelés par les règles sont les suivants les niveaux :

5.1. Lexical

- Mots dont l'initiale "ai-" se prononce [e] (et non pas [ɛ]), (25 mots) "aider, aigri"...
- Mots dont la finale "-c" est muette, (20 mots) "accroc, ajonc, banc, blanc, broc"...
- Mots dont la finale "-d" se prononce (mots d'emprunt), (42 mots) "apartheid, background"...
- Mots contenant un "-e-" prononcé [e], (52 mots) "siderata, referendum, revolver"...
- Mots dont la finale "-et" se prononce [ɛt], (16 mots) "basket, cricket, gadget"...
- Mots dont la finale "-ey" se prononce [ɛ], (13 mots)"jockey, bey"...
- Mots contenant "-gu-" (avant "i") prononcé [g] (7 mots) "anguille, figuier, déguiser"...
- Mots dont la finale "-g" ne se prononce pas, (24 mots) "barlong, bastaing, basting, bourg"...
- Mots dont l'initiale "h-" est aspirée[·], (403 mots) "ha, habanera, hâblerie, hâbleur"...
- Mots dont la finale "-ing" se prononce [iŋ], (66 mots) "antifading, bowling, brainstorming"...
- Mots dont la finale "-l" ne se prononce pas, (22 mots) "cul, cucul, outil"...
- Mots qui contiennent "-oo-" prononcé [u], (18 mots) "hollywood, lambswool"...
- Mots dont la finale "-p" ne se prononce pas, (15 mots) "beaucoup, camp, cantaloup"...
- Mots dont l'initiale "qu-" se prononce [kw], (53 mots) "quadragénaire, quadragésimal"...
- Mots dont la finale "-s" se prononce [s], (510 mots) "abribus, acarus, acinus"...
- Mots dont la finale "-t" se prononce [t], (194 mots)"abject, abrupt, accessit, achromat"...
- Mots qui contiennent "-un-" prononcé [ɔ̃], (22 mots) "acupuncture, avunculaire"...
- Mots dont la finale "-x" ne se prononce pas, (44 mots) "afflux, affûtiaux, alquifoux"...
- Mots dont l'initiale "gn-" se prononce [ɲ], (8 mots) "gnon, gniaf"...
- Mots dont la finale "-er" se prononce [ɛʀ] , (86 mots) "africaner, hamster, container, revolver,"...
- Mots dont la finale "-er" se prononce [œʀ], (33 mots) "angledozer, bitter, bookmaker, booster"...

5.2. Morphophonémique

- Préfixes qui finissent par une voyelle et qui entraînent la transcription de "s" suivant en [s] : "a, aéro, anti, auto, bi, carbo, entre, extra, para"... "asocial, aérosol, antisocial, autosensoriel"...
- Bases préfixales (mots grecs) qui finissent par une voyelle et qui entraînent la transcription de "s" en [s] : "éco, immuno, hypo, homo, thio, pyro"... "écosystème, immunosuppresseur"...
- Préfixes qui finissent par "y" et qui entraînent la transcription de "ya" en [ia] : "poly, oxy, tachy" "polyacide, oxyacétylène, tachyarythmie"...
- Préfixes qui finissent par "i" et qui entraînent la transcription de "a" en [ia] (et non pas [ija]) : "di, bi, tri, anti, vivi, uni, maxi, mini" "diacétone, biannuel, triannuel, antiarmement"...

5.3. Catégoriel (flexionnel)

Mots dont la finale "-tions" se prononce [siɔ̃], c'est-à-dire tous les noms (*i.e.* les "non-verbes") contient 1 780 mots (abdications, abductions, aberrations, abjections...). C'est pourquoi, nous avons conservé le lexique contraposé, c'est-à-dire le lexique des verbes.

- Formes verbales dont la finale est "-tions" (prononcée [tiɔ̃], *i.e.* tous les verbes en "-tions" sauf "argutions, balbutions, initions"), (825 mots)
  (verbes en "-tir") mentions, partions...
  (verbes en "-ter") abritions, absentions...
- Mots dont la finale "-ent" se prononce [ɑ̃], c'est-à-dire tous les "non-verbes", plus ("sent..."), (2 680 mots) "abaissement, abâtardissement"...

5.4. Syntaxique

Ces listes sont données à titre informatif : un lexique de ces formes intégré à TOPH ne peut lever les ambiguïtés qui sont d'ordre syntaxique.

- Homographes dont la finale "-tions" se prononce [tiɔ̃] pour la catégorie verbale et se prononce [siɔ̃] pour la catégorie des noms : "ablations, acceptions, adoptions, affections"...
- Homographes dont la finale en "-ent" ne se prononce pas pour la catégorie verbale et se prononce [ɑ̃] pour les autres catégories : "coïncident, content, couvent, divergent"...

6. CONCLUSION

La nouvelle grammaire a permis la phonétisation du *60 000*. Les formes phonétiques sont correctes avec un score avoisinant 100%. Les erreurs résiduelles sont dues à des cas particuliers et qui ne peuvent pas être traités par une grammaire de type TOPH.

Nous avons voulu mettre en évidence, dans cette étude, le rôle prépondérant que jouent les lexiques intégrés dans la grammaire. Ils ne doivent pas être considérés comme de simples collections de données mais au contraire comme des représentations naturelles dans la modélisation conjointe lexique-règles que propose TOPH.

Cette étude s'est attachée à la description la plus approfondie possible des problèmes de la phonétisation du mot. Nous l'avons poursuivie par l'étude des mots composés [Belrhali *et al.* 1991]. Une seconde étape de ce travail va consister à définir et introduire dans *Le 60 000* un ensemble de catégories lexicales spécifiques aux phénomènes de l'oral (désambiguïsation de la phonétisation et de l'intonation). En adoptant et adaptant un modèle flexionnel [Rouault, 1988] ces catégories vont permettre de générer l'ensemble des formes déduites (environ 450 000) en ajoutant à TOPH un champ pour la catégorie [Chatti, 1991]. Une nouvelle grammaire sera établie à partir de ce lexique de formes, et sera complétée par l'étude des variantes en nous appuyant sur BDPHO, une base de données du français parlé [Boë et Tubach, 1992].

RÉFÉRENCES

Aubergé V., *Contribution à la phonétisation automatique des langues alphabétiques : le langage TOPH*, Rapport de DEA, CRISS, Univ. Grenoble II, 1985.

Aubergé V., *La synthèse de la parole : "des règles aux lexiques"*, Thèse de Doctorat, Univ. P. Mendès France, Grenoble, 1991.

Aubergé V., Contini M., Maret D. & Schnabel B., *TOPH : un outil de phonétisation multilangue*, Bull. de l'Inst. de Phon. de Grenoble, 16, 155-176, 1987.

Belrhali R. & Libert L., *Phonétisation automatique : évaluation et enrichissement d'une grammaire de phonétisation du français*, Mémoire de DEA, Univ. Stendhal, Grenoble, 1991.

Berlrhali R., Libert L., Aubergé V. & Boë L.J., *Organisation de lexiques pour une grammaire de phonétisation du français : "Le 60 000" de l'ICP.*, Séminaire Lexique, GDR-PRC Com. HM, Toulouse, 1992.

Boë L.J. & Tubach J.P., *BDPHO : une base de données lexicales orthographique-phonétique du français parlé.*, Séminaire Lexique, GDR-PRC Com. HM, Toulouse, 1992.

Catach N., *La phonétisation automatique du français*, Ed. CNRS, Paris, 1984.

Chatti C., *Catégories lexicales du français*, Mémoire de Maitrise, Univ. Stendhal, Grenoble, 1991.

Divay M. & Guyomard M., *Conception et réalisation sur ordinateur d'un programme de transcription graphémo-phonétique du français*, Thèse de 3ème cycle, Univ. de Rennes, 1977.

Grammont M., *Traité pratique de prononciation française*, Ed. Delagrave, Paris, 241 p., 1926.

Juilland A., *Dictionnaire phonétique inverse du français.*, Ed. Mouton, The Hague, Paris, 1965.

Lacheret-Dujour A., *Contribution à l'analyse de la variabilité phonologique pour le traitement automatique de la parole continue multilocuteur*, Thèse de Doctorat, Univ. de Paris 7, 1990.

Lahens F., *Un modèle stochastique pour la vérification et la correction automatique de textes : le système VORTEX*, Thèse de 3ème cycle en informatique, CERFIA, Toulouse, 1987.

Laporte E., *Méthodes algorithmiques et lexicales de phonétisation de textes, applications au français*, Thèse de 3ème cycle, Univ. Paris VII, 161 p., 1988.

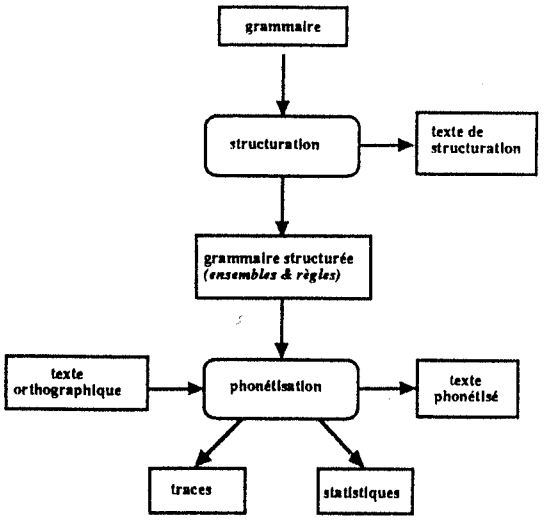
Prouts B., *Contribution à la synthèse de la parole à partir du texte, transcription graphème-phonème en temps réel sur microprocesseur*, Thèse de Docteur Ingénieur, Univ. Paris Sud-Orsay, 1980.

Rey-Debove J., *Le Robert Oral-Écrit, l'orthographe par la phonétique*, Le Robert, 1 400 p., 1989.

Robert P., *Le Petit Robert 1*, Société du Nouveau Littré, Le Robert, 1969 p., 1990.

Rouault J., *Linguistique automatique, Applications documentaires*, Sciences pour la Communication, Peter Lang Ed., Berne, 309 p., 1987.

Figure 1 : TOPH : une organisation modulaire



ANNEXE
Un extrait de la phonétisation : déclarations générales et exemples pour la classe du graphème "a".
(langage **TOPH**, version 2.1)

• Déclarations

Ensembles

Linguistiques
"Lettres" = (a, à, â, b, c, ç, d, e, é, è, ê, ë, f, g, h, i, î, ï, j, k, l, m, n, o, ô, p, q, r, s, t, u, ù, ü, û, v, w, x, y, z)
"Voy." = (a, à, â, e, é, è, ê, ë, i, î, ï, o, ô, u, ù, ü, û, y)
"E & I & Y" = (e, é, è, ê, ë, i, î, ï, y)
"A & O & U" = (a, à, â, o, ô, u, ù, ü, û)
"Consonnes" =
(b, c, ç, d, f, g, h, j, k, l, m, n, p, q, r, s, t, v, w, x, z)
"Consonnes doubles" =
(bb, cc, dd, ff, gg, ll, mm, nn, pp, rr, ss, tt)
"Consonnes liquides" = (l, r)
...

Para-linguistiques
"Chiffres" = (0, 1, 2, 3, 4, 5, 6, 7, 8, 9)

Frontière de mots
"#" = (/(, /), -, / , /, , ., :, ;, !, ?, ", ')

Lexique
"Mots pour lesquels "ai "est prononcé [e]"
"aider, aigri, aigrir, aigu, aiguillage"...

• Règles

Emprunts anglais
("#"+ch,m,p) +a+ (ce,de,ker,se) = [ɛ]
steeple-chase, pace-maker, self-made-man
("#"+b,c,footb,w) +a+ (ll,lk) = [o]
ball, call-girl, football, walk-over
...

Emprunts allemands
("#"+schn) +au+ (zer) = [aw] *schnauzer*
("#"+sprechges) +ang+ = [aŋ] *sprechgesang*
...

Emprunts italiens
("#"+qu) +an+ (ti) = [an] *tutti quanti*
...

Emprunts espagnols
("#"+c) +au+ (dillo) = [aw] *caudillo*
+a+ (yuntamiento) = [a] *ayuntamiento*
...

autres
("#"+grœnend) +ae+ (l) = [a] *flamand*
...

• Règles générales

+am+ (h,s)	= [am]	*amharique*
+am+ ("CONS. NON NAS.","#")	= [ɑ̃]	*amphibie*
("#") +an+ (h)	= [an]	*anhydride*
+an+ ("#","CONS. NON NAS.",m)	= [ɑ̃]	*manger*

...

("#"+f,p,t) +aon+	= [ɑ̃]	*paon*
+aim+ ("#")	= [ɛ̃]	*daim*
+ain+ ("#","CONS. NON NAS.")	= [ɛ̃]	*pain*
+a+ (il+l,"#")	= [a]	*travail*
("#") +aî+ (né)	= [e]	*aîné*

...

("#"+abb) +aye+	= [eĭ]	*abbaye*
(p) +ay+ (s)	= [eĭ]	*pays*
("#"+f) +a+ (ya)	= [a]	*fayard*
("#"+b,c,cob,cong,rim) + a+ (yé)	= [a]	*bayer*
+a+ (y+é,er,iste)	= [e]	*payer*
("#"+f,m) +a+ (yo)	= [a]	*fayot*
+a+ (yure)	= [e]	*balayures*
+a+ (y+"VOYELLE")	= [ɛ]	*maye*
(f) +ai+ (san)	= [ə]	*faisan*
("#"+qu,g) +ai+ ("#")	= [ɛ]	*quai*
("#") +ai+ ("EXCEP:init.enAI")	= [e]	*aider*
("#"+g,s,p) +aie+ (tt,rie,"#")	= [e]	*paierie*

...

	+ai+		= [ɛ]
	+âi+		= [a]
	+aî+		= [ɛ]
	+a+	(ï)	= [a]
	+æ+		= [e]
	+à+		= [a]
	+â+		= [a]
	+au+		= [o]
	+ay+	("#")	= [ɛ]
	+aa+		= [a]
("#")	+a+	("#")	= [a]
	+a+		= [a]

19es J.E.P. - Bruxelles, 19 au 22 mai 1992.

LES FRICATIVES DE L'ARABE SOUS *SONIA*

ABDELKADER BETARI - REMY BULOT

G . I . A - URA 816

Résumé

L'arabe se distingue des langues indo-européennes par son caractère emphatique et fricatif. En raison de la place primordiale des consonnes constrictives dans le système phonétique de l'arabe, nous avons décidé d'analyser en premier ces phonèmes dans le but de les reconnaître automatiquement dans le discours continu. La reconnaissance de ces phonèmes est entièrement effectuée sous *SonIA* qui est un environnement PrologII enrichi de prédicats évaluables pour la paramétrisation du signal et la reconnaissance de formes.

Pour réaliser la partie "ascendante" du Décodage Acoustico-Phonétique, on utilise un ensemble de références spectrales correspondant à la tenue des constrictives. L'étude acoustique de ces phonèmes nous a conduit à considérer trois catégories importantes pour la reconnaissance de ces phonèmes : les fricatives qui peuvent être emphatiques, les fricatives avants qui sont peu sensibles au contexte phonétique et les fricatives arrières pour lesquelles les effets de coarticulation sont tels que l'identification de ces consonnes ne peut se faire correctement hors contexte.

1. INTRODUCTION

Depuis une trentaine d'années surtout, la vive renaissance de la vie intellectuelle dans les pays de langue arabe a réclamé une évolution de la langue vers des moyens propres à exprimer des idées modernes et des sentiments nouveaux. Renonçant avec raison à les emprunter aux langues européennes, l'arabe cherche son enrichissement dans le développement des ressources naturelles de la langue. Pour notre part, nous nous intéressons plus particulièrement à la Reconnaissance Automatique de l'arabe parlé et à son étude dans les 3 domaines : phonétique, articulatoire et acoustique.

La langue dont nous privilégions l'analyse est l'arabe classique, tel qu'elle est pratiquée actuellement, appelée aussi "arabe littéraire" ou "arabe standard contemporain". Cette langue présente la particularité d'être parlée et écrite par une grande partie des populations arabes et musulmanes appartenant aux milieux cultivés. Elle se distingue de l'*arabe dialectal* (ou vulgaire), représenté par la multitude de parlers dans les différents pays arabes.

L'arabe se distingue des langues indo-européennes par son caractère emphatique et fricatif. En effet son alphabet (qui compte vingt-huit lettres) contient pour moitié des consonnes constrictives. Aussi, en raison de leur place primordiale dans le système phonétique de l'arabe, nous avons décidé d'analyser en premier ces consonnes constrictives dans le but de les reconnaître automatiquement dans le discours continu. La reconnaissance de ces phonèmes est entièrement effectuée sous *SONIA* qui est un environnement PrologII enrichi de prédicats évaluables pour le traitement du signal et la reconnaissance de formes.

Bien que la reconnaissance fonctionne en mono-locuteur, l'adaptation à un nouveau locuteur ne pose pas de problème particulier.

2. REPRESENTATION PARAMETRIQUE ET OUTILS

Le signal de parole est numérisé sur 16 bits à une fréquence de 16 kHz puis préaccentué et caractérisé chaque 10 ms par son énergie globale, la densité des passages par zéro et les énergies spectrales dans 24 canaux répartis suivant une échelle de Mel (figure 1 et 2). Les spectres peuvent être obtenus par différentes méthodes (FFT, LPC, Cepstre, Vocodeur, modèle d'oreille). Le choix de cette représentation est dicté par l'impératif de suivre au plus près la formalisation des connaissances acoustiques et phonétiques proposée par les experts de ce domaine.

Un ensemble d'outils permet de définir et de calculer dynamiquement sous Prolog de nombreux

paramètres auxiliaires obtenus par combinaisons des 26 attributs initiaux [Méloni 86], [Bulot 87]. Dans le cadre du système de DAP "ascendant" nous utilisons essentiellement comme paramètres secondaires les valeurs de différentes distances [Applebaum 87] [Itakura 87] entre les spectres de référence d'un locuteur et ceux calculés sur le signal, diverses fonctions d'instabilité spectrale ainsi que quelques indices évaluant contextuellement les caractères vocalique et consonantique.

La localisation de certaines zones intéressantes sur les courbes caractérisant l'évolution temporelle des paramètres est effectuée au moyen de prédicats évaluables qui permettent de définir des schémas de formes (pics, vallées, segments monotones, etc.) et d'identifier les événements correspondant à certaines de ces descriptions. De plus, ces outils assurent le passage de la représentation numérique du signal vers divers types d'unités symboliques plus ou moins complexes constituées d'associations de formes simples [Méloni 86], [Bulot 87].

Cet environnement de travail pour la représentation et le traitement de connaissances acoustiques, phonétiques et linguistiques a été utilisé pour réaliser divers systèmes de DAP et de reconnaissance de la parole [Bulot 87] [Méloni 91]. Les performances des techniques utilisant des connaissances explicites sont liées à la quantité d'informations symboliques "sûres" disponibles lors de l'activation des règles. Ces méthodes conviennent donc mieux à une phase "descendante" du processus de reconnaissance. Ces difficultés nous ont conduit à la réalisation du système actuel de DAP "ascendant" auquel sera associé un processus de vérification descendante sur les cohortes de mots déduites du treillis phonétique.

3. STRATEGIE GENERALE DE RECONNAISSANCE

Pour réaliser la partie "ascendante" du Décodage Acoustico-Phonétique, on utilise un ensemble de références spectrales correspondant aux phases stables des phonèmes (tenue des constrictives). Le but ici n'est pas de décrire précisément les sons - qui de toute manière sont plus ou moins fortement déformés par le contexte - mais de proposer une forme moyenne proche de la réalisation idéale du phonème.

L'acquisition des références est effectuée à partir d'un ensemble très limité de phrases dans lesquelles les phonèmes apparaissent soit dans des contextes peu déformants, soit au contraire dans des contextes très particuliers. Cette phase d'adaptation est donc assez rapide et peu contraignante.

Différents types de distance aux références sont systématiquement calculés et constituent les paramètres essentiels pour localiser et identifier les phonèmes. Les intervalles du signal dont l'une des distances est inférieure à un seuil très large sont potentiellement considérés comme candidats probables. Nous verrons que les décisions sont d'autant plus faciles à prendre que les algorithmes de calcul des distances intègrent une certaine "intelligence" résultant de la prise en compte de connaissances diverses.

Un ensemble très restreint de règles simples utilise les hypothèses déduites à partir des distances ainsi que certains paramètres supplémentaires (instabilités spectrales, caractère vocalique ou consonantique, etc.) pour proposer des unités phonétiques. La segmentation ainsi obtenue de l'énoncé n'est pas déterministe et les frontières entre les unités phonétiques ne sont pas strictes.

La variabilité acoustique des phonèmes - résultant essentiellement des phénomènes de coarticulation - est importante et interdit une identification très précise des unités phonétiques. Cependant, les distances doivent mettre en valeur les informations les plus pertinentes dans des contextes divers. Compte tenu du système de représentation paramétrique que nous avons choisi, notre travail a consisté à résoudre les problèmes suivants :

- adaptation au locuteur rapide et peu contraignante,
- ajustement des niveaux d'énergie spectrale,
- prise en compte optimale des variations de position et d'amplitude des maxima spectraux (formants),
- intégration d'informations contextuelles disponibles dans le signal.

4. LES FRICATIVES DE L'ARABE

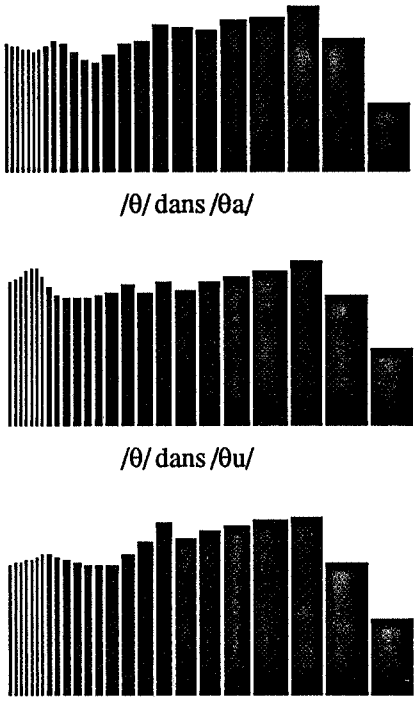
Les consonnes fricatives de l'arabe sont des sons résultant d'un resserrement ou constriction en un ou plusieurs points du conduit vocal. Il est possible de produire des consonnes fricatives en n'importe quelle partie de l'appareil vocal depuis les lèvres jusqu'au pharynx et même au larynx (c'est le cas du son /h/) [Al-ani 70]

Les fricatives sont classées dans le plan articulatoire selon la zone où se fait la constriction maximale. De l'avant vers l'arrière du conduit vocale on rencontre : /f/ (labiale), /θ/, /ð/ (interdentales), /s/, /z/ (dento-alvéolaires), /ʃ/, /ʒ/ (prépalatales), /χ/, /ɣ/ (postlvélaires), /ʕ/, /ħ/ (pharyngales), et /h/ (laryngale).

D'un point de vue acoustique, elles se traduisent par la présence d'un bruit constitué de vibrations apériodiques étalées sur le spectre. En outre, on observe parfois sur l'axe temporel de véritables petits silences (10-20 ms) appelés closions qui encadrent parfois les fricatives sourdes [Calliope 89] [Soussi 81].

L'étude acoustique des fricatives de l'arabe nous a rapidement conduit à considérer deux classes importantes pour la reconnaissance de ces phonèmes :

- d'une part les labiales, dentales et palatales (/f/, /ð/, /θ/, /s/, /z/, /ʃ/, /ʒ/) dont le spectre est peu sensible au contexte phonétique (leur reconnaissance pourra se faire indépendamment du contexte, figure 1),
- d'autre part les vélaires, pharyngales et laryngale (/χ/, /ɣ/, /ʕ/, /ħ/, /h/)dont le point de constriction très arrière dans le conduit vocal permet une anticipation ou une prolongation articulatoire importante des phonèmes voisins. Les déformations acoustiques qui en résultent sont généralement importantes [Soussi 81] [Ghazali 87] (figure 2) et l'identification de ces consonnes ne peut se faire correctement hors contexte.



/θ/ dans /θa/

/θ/ dans /θu/

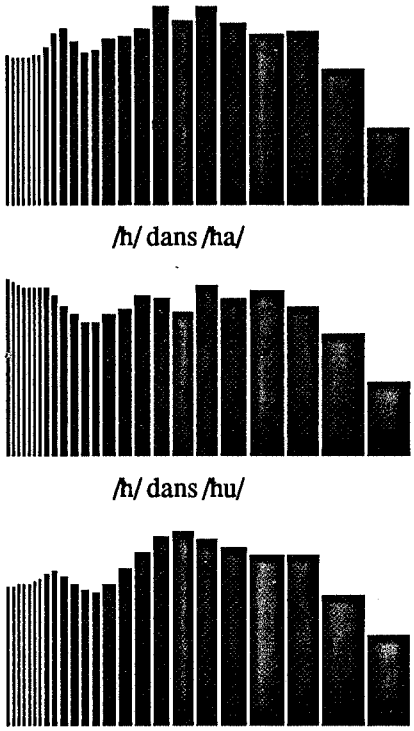
/θ/ dans /θi/

Figure 1 : Spectres de /θ/ dans différents contextes vocaliques.

Dans le processus de reconnaissance, différentes distances aux références sont systématiquement calculées et constituent les paramètres essentiels pour construire des paramètres phonétiques qui sont des courbes temporelles de proximité dont les valeurs varient entre 0 et 100.

Les fricatives sont localisées et identifiées à partir de schémas de formes (généralement des collines) émergeant suffisamment sur ces courbes. Chaque segment retenu reçoit une étiquette (liée à la référence sur laquelle a été construite la courbe de proximité atteinte sur le segment) ainsi qu'une valuation (valeur maximum de la courbe de proximité qui a permis de détecter le segment). Toutefois un seuil minimum est exigé pour cette valuation pour que l'événement détecté soit considéré comme pertinent. Ces segments sont ensuite ajoutés dans un treillis de résultats ; ils seront utilisés ultérieurement par les niveaux supérieurs (lexique, syntaxe, sémantique, etc.) pour la reconstruction du message vocal.

Toutefois, cette caractérisation des sons fricatifs reste insuffisante puisque ces consonnes peuvent-être éventuellement géminées (allongement de la durée). En plus, /ð/ et /s/ peuvent-être emphatique (/ð̰/ et /s̰/), ce qui se traduit par une compacification du spectre de ces consonnes ainsi que de la voyelle de la syllabe concernée [Bonnot 77]. Ces informations sont pertinentes pour les accès lexicaux et doivent permettre de limiter les mots candidats.



/ħ/ dans /ħa/

/ħ/ dans /ħu/

/ħ/ dans /ħi/

Figure 2 : Spectres de /ħ/ dans différents contextes vocaliques.

5. RECONNAISSANCE DE /f/, /θ/, /z/, /ʃ/, /ʒ/

Les 5 phonèmes /f/, /θ/, /z/, /ʃ/, /ʒ/ sont reconnus séparément et de manière indépendante. Aussi nous prendrons la consonne /θ/ comme exemple sachant que les autres fricatives de cette catégorie sont reconnues suivant le même procédé moyennant quelques variantes sur la fonction de distance utilisée.

Nous disposons au départ d'un spectre de référence choisi dans la phase stable de /θ/ et dans un contexte peu déformant. La reconnaissance est faite hors contexte et nous procédons de la manière suivante :
- Nous construisons sur toute la phrase une courbe de proximité de cette référence à l'aide d'une distance sur la dérivée du spectre (voir figure 3).

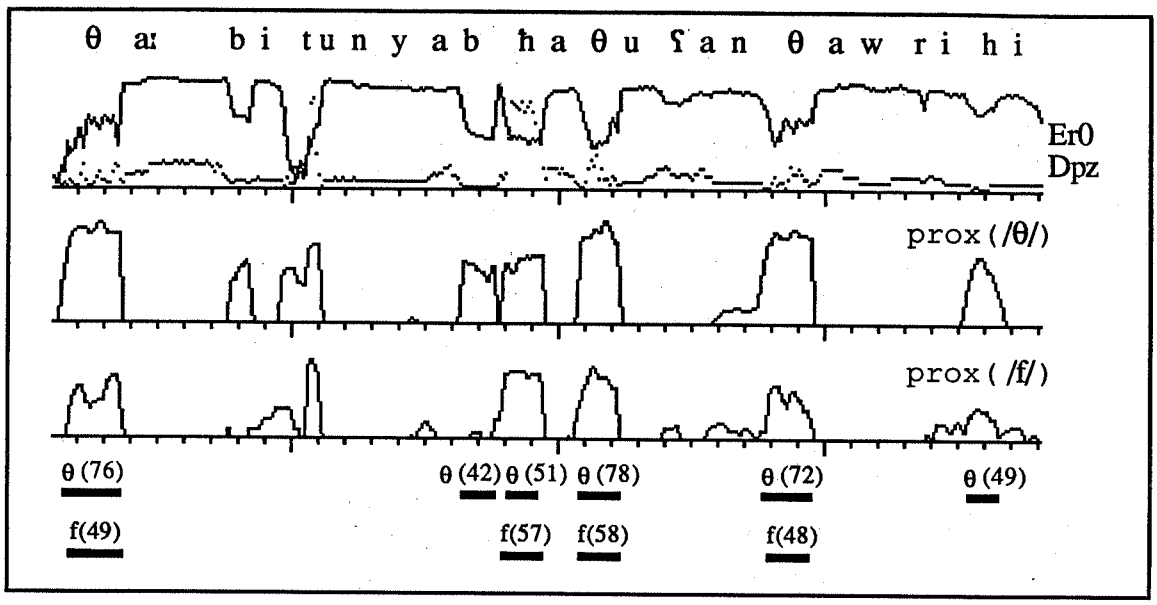


figure 3: Courbes de proximité sur /θ/, /f/ et segments phonétiques détectés avec leurs scores.

- Cette courbe est lissée par un filtre passe-bas choisi pour éliminer les pics erratiques de durée trop faible pour constituer des événements phonétiques en eux mêmes.
- Sur cette courbe ainsi prétraitée qui évolue dans l'intervalle [0,100], nous recherchons toutes les collines émergeant à droite et à gauche d'au moins 20 points et ayant un maximum relatif supérieur à 40 points (seuil minimum de pertinence). Cette forme est définie par le prédicat Prolog:

```
colline_fric(p,z,s) ->
   colline(p,z,40,20,20,s);
```

où p est le paramètre phonétique analysé, z la portion de signal explorée et s le segment résultat sur lequel est observé la colline. Les segments s détectés dans z sont énumérés par backtracking.

Ce schéma de forme est utilisée dans la régle:

```
segment_fric(z,<"θ",s,v,g>) ->
   colline_fric(Prox("θ"),z,s)
   inferieur(5,longueur(s))
   valeur_sup(Prox("θ"),s,v)
   coef_gemination("θ",s,g) ;
```

Un coefficient g de gémination est calculé en fonction de la durée du segment et du phonème ; cette information permettra d'orienter en partie la recherche des mots dans le lexique. Enfin, un taux de vraisemblance v (directement déduit de l'émergence de la colline) est attribué au segment ainsi étiqueté. La règle `segment_fric` énumère par backtracking tous les segments de la phrase vérifiant les contraintes demandées. Les résultats sont ajoutés dans le treillis phonétique sous la forme d'un terme:

`<"θ",<début,fin>,score,gémination>` (1)

6. LES EMPHATIQUES

Les deux phonèmes /s/, /ð/ peuvent-être emphatiques suivant le mot dans lequel ils apparaissent et il en résulte sur le plan acoustique une déformation non négligeable du spectre (compacification). Aussi, une seule référence spectrale ne peut correctement convenir à la fois pour une situation emphatique et pour une situation non emphatique. Nous avons remédié à ce problème en sélectionnant 2 références pour chacun de ces 2 phonèmes : une référence emphatique et une référence non emphatique.

Sur le plan de la stratégie, les réalisations emphatiques /s̲/, /ð̲/ sont considérées comme 2 fricatives supplémentaires dans le système phonétique et sont recherchées dans le signal de la même manière que les autres phonèmes.

Sur le plan des résultats, les étiquettes emphatiques et non emphatiques sont souvent concurrentes sur le même segment avec des scores importants, mais la bonne étiquette est la plus part du temps classée en première position (figure 4).

Nous savons aussi que le caractère emphatique des consonnes se prolonge dans la voyelle associée. Aussi cette propriété devra être vérifiée par les niveaux supérieurs lorsque les accès lexicaux permettront dereconstruire la syllabe. Cette contrainte est représentée

1 Les mots en italique dans les termes Prolog désignent des résultats (généralement des constantes entières) qui n'ont pas de sens hors application.

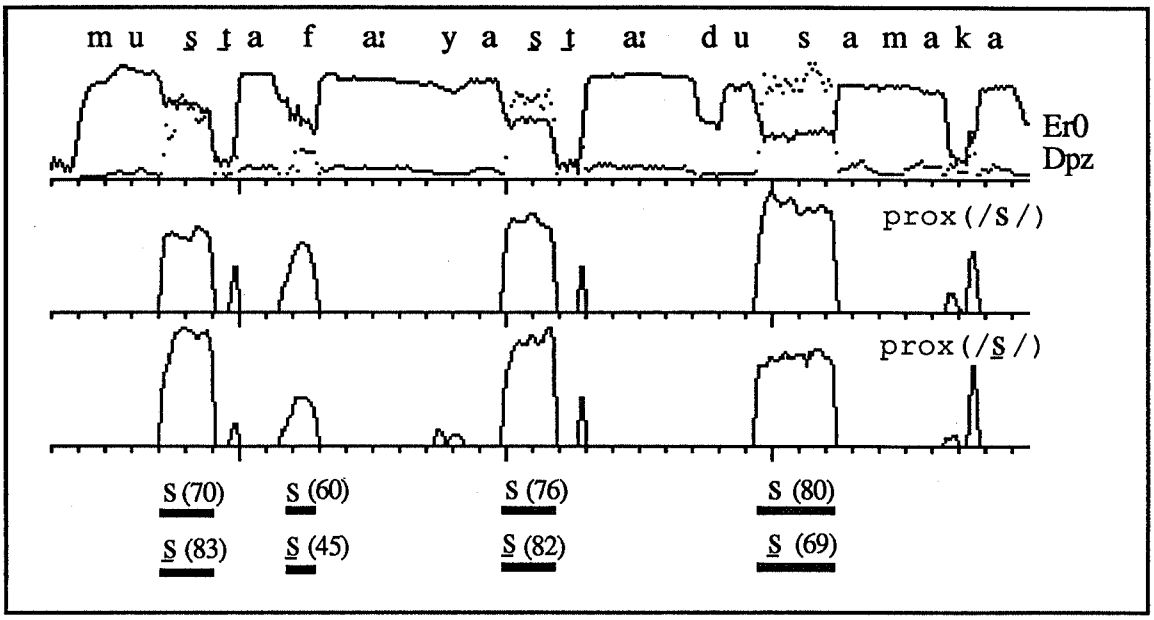


figure 4 : Courbes de proximité sur /s/, /s̪/ et segments détectés avec leurs scores.

par un ajout conditionnel dans le treillis de résultats :

```
si_condition(<"s̪", seg, score, gém>,
             voyelle_associee(v).
             emphatique(v).nil)
```

7. LES FRICATIVES ARRIERES

Les fricatives /χ/, /ɣ/, /ʕ/, /ħ/, /h/ articulées en arrière du conduit vocal ont une forme fortement colorée par les voyelles environnantes. Les déformations sont telles que nous avons décidé de choisir une référence spectrale pour chaque contexte vocalique différent.

Pour chacune de ces fricatives nous avons donc 3 références distinctes choisies dans les contextes vocaliques /a/, /u/, /i/ (le système phonétique de l'arabe ne contient que ces trois voyelles). La stratégie de reconnaissance reste identique à celle décrite précédemment et chaque référence sert à détecter ces fricatives dans les contextes vocaliques pour lesquels elles ont été choisies. Aussi le résultat n'a de sens que si le contexte vocalique fixé à priori est bien vérifié par la suite.

Cette vérification sera demandée au niveau supérieur par l'intermédiaire du treillis de résultats dans lequel on effectue un ajout conditionnel (figure 5). Si on suppose par exemple qu'un segment *s* a été détecté grâce à la référence /χa/ (spectre de /χ/ en contexte /a/), alors le résultat suivant est ajouté :

```
si_condition(<"χ", s, score, gémination,
             voyelle_associee("a"))
```

Lors des accès lexicaux, cette unité phonétique (défini par un intervalle de temps, une étiquette "χ" et un score) ne pourra apparaître que dans des syllabes contenant la voyelle /a/ (figure 5).

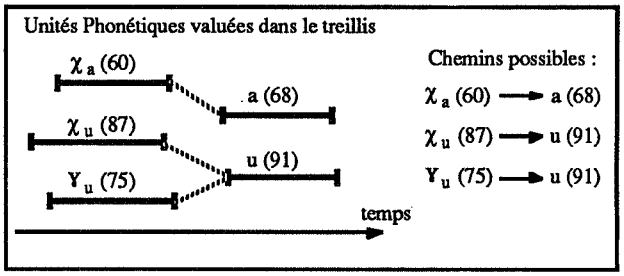


Figure 5 : exemple de contraintes sur des combinaisons d'unités phonétiques dans le treillis.

Bien qu'assez sommaire et simpliste, ces contraintes sur le contexte phonétique (limitées à la nature de la voyelle contenue dans la syllabe) améliore nettement les performances de reconnaissance des fricatives arrières.

8. RESULTATS ET DISCUSSION

Nous nous sommes servis d'un corpus de 130 phrases phonétiquement équilibrées pour l'arabe et toutes les séquences "consonne-voyelle", "voyelle-consonne" sont présentes aux moins une fois. Les performances de reconnaissance des fricatives sont résumées dans la matrice de confusion ci-dessous : il s'agit de taux de réussite sur la première étiquette proposée pour chaque phonème. La colonne *"1&2"* cumule les fois ou la bonne étiquette apparaît dans les deux premiers candidats.

%	/f/	/θ/	/s/	/ʃ/	/ð/	/z/	/ʒ/	/χ/	/ɣ/	/ʕ/	/ħ/	/h/	1&2	Echec
/f/	64	20	0	3	6	0	5	0	2	0	0	0	85	3
/θ/	9	79	0	0	10	0	0	1	1	0	0	0	90	2
/s/	3	3	90	0	2	0	2	0	0	0	0	0	95	2
/ʃ/	4	2	0	91	0	1	2	0	0	0	0	0	94	1
/ð/	2	1	0	0	91	0	0	1	5	0	0	0	94	1
/z/	0	0	3	0	0	89	5	0	3	0	0	0	93	2
/ʒ/	2	3	0	9	0	2	84	0	0	0	0	0	89	0
/χ/	5	4	0	0	1	0	0	80	7	0	0	3	90	1
/ɣ/	4	0	0	0	0	0	0	6	84	4	2	0	91	2
/ʕ/	0	0	0	0	1	0	3	2	5	80	5	4	88	5
/ħ/	2	1	0	0	0	0	0	1	1	10	78	7	89	2
/h/	1	2	0	0	0	0	1	4	3	6	4	79	85	4

Matrice de confusion sur le premier candidat en pourcentage.

La colonne "*Echec*" comptabilise le taux de fricatives non détectées (non localisées ou scores trop faibles).

Bien que déjà satisfaisants pour une analyse remontante, les résultats pourront être améliorés en tenant compte du contexte acoustique (certaines contraintes contextuelles sur le signal devraient permettre d'affiner les distinctions entre des phonèmes acoustiquement proches et éliminer des cas d'assimilation. Ex: /f/ - /θ/ ou /χ/ - /ɣ/).

Le score attribué aux phonèmes pourrait être épaulé par d'autres facteurs (la base de son est en cours d'extension, on pourra étalonner nos paramètres phonétiques en tenant compte des taux de réussite en fonction des valeurs de proximité et construire ainsi une valuation statistique moins arbitraire que celle actuellement utilisée).

9. CONCLUSION

L'environnement Prolog que nous utilisons facilite l'expertise acoustico-phonétique et nous a permis de créer rapidement un prototype pour valider nos connaissances. Bien que nous ayons restreint l'application à la reconnaissance mono-locuteur, l'adaptation à un nouveau locuteur s'est révélée particulièrement aisée puisqu'elle consiste à choisir un nombre restreint de références spectrales. Toutefois, cette opération nécessite quand même l'intervention d'un expert car la qualité des références conditionnent largement les performances de reconnaissance.

Bien que perfectible, cette phase ascendante du DAP fournit des résultats très satisfaisants, aussi bien du point de vue de la localisation des fricatives que du point de vue de leur identification. La reconnaissance des autres phonèmes est en cours de développement et permettra bientôt de fournir des treillis phonétiques complet. La phase d'analyse descendante sera prochainement abordée : elle devrait permettre de limiter et de classer précisément les solutions concurrentes dans un contexte connu.

10. BIBLIOGAPHIE

[Al-ani 70] Al-ani S. H. : *Arabic Phonology, An acoustical and Physiological Investigation. Mouton* ; The Hague Paris 1970.

[Applebaum 87] Applebaum T.H., Hanson A.H., Wakita H. : *Weighted distance measures in vector quantization based speech recognizers* ; Proceedings ICASSP, pp. 1155-1158.

[Bonnot 77] Bonnot : *Recherche expérimentale sur la nature des consonnes emphatiques de l'arabe classique* ; rapport n°9 de l'institut phonétique de Strasbourg, 1977

[Bulot 87] Bulot R. : *Techniques d'Intelligence Artificielle pour la reconnaissance de la parole : application au décodage acoustico-phonétique* ; Thèse de l'université d'Aix-Marseille II.

[Calliope 89] : *La parole et son traitement automatique* ; Masson 1989.

[Ghazali 87] Ghazali S. : *Etude EMG préliminaire sur les consonnes arrières de l'arabe*, 16eme J.E.P Hammamet-Tunisie, 1987. P 286- 289.

[Itakura 87] Itakura F., Umesaki T. : *Distance measure for speech recognition based on the smoothed group delay spectrum* ; Proceedings ICASSP, Dallas, pp. 1257-1260.

[Méloni 86] Meloni H., Bulot R. : *Un système de traitement de connaissances pour le décodage acoustico-phonétique* ; ICA Symposium on speech recognition, Mc Gill University.

[Méloni 91] Meloni H., Gilles P., Betari A. : *Representation of acoustic and phonetic knowledge for speaker-independent recognition of small vocabularies* ; Speech Communication, North Holland, Amsterdam, volume 10, n° 2, June, 1991.

[Soussi 81] Soussi I. : *Analyse acoustique et phonétique des consonnes de l'arabe dans le parler de Damas* , Thèse de troisième cycle 1981, Université de Provence, 1981.

19es J.E.P. - Bruxelles, 19 au 22 mai 1992.

RESTRICTIONS SEMANTIQUES APPORTEES A L'ETUDE DES GROUPES NOMINAUX EN 'NDEN' : APPLICATION A LA MACHINE A DICTER

J. KLEIN, K. SMAILI, L. ROMARY, F. CHARPILLET

CRIN INRIA LORRAINE
BP 239 54506 VANDŒUVRE-LES-NANCY

Résumé

Dans cet article, nous présentons comment l'intégration d'un analyseur sémantique permet d'améliorer les résultats de la machine à dicter. L'analyseur sémantique ne traite actuellement que des expressions de la forme *Nom* de *Nom*. La machine à dicter quant à elle fonctionne à l'aide de modèles biclasses et triclasses essentiellement syntaxiques et, vu les incertitudes de la reconnaissance de la parole, fournit en résultat un ensemble de phrases syntaxiquement correctes mais qui peuvent être totalement dépourvues de sens. l'analyseur sémantique permet de sélectionner parmi les expressions en NdeN celles qui sont porteuses de sens et ainsi de diminuer le nombre de solutions possibles, rendant l'outil plus performant et donc plus ergonomique.

INTRODUCTION

Les systèmes de traitement automatique du langage naturel peuvent être classés en deux grandes catégories : les systèmes syntaxiques tels que la machine à dicter, où la compréhension n'est pas nécessaire, et les systèmes où la compréhension, donc l'analyse sémantique, est indispensable pour accomplir une tâche par exemple comme dans les systèmes de commande, de dialogue ou d'interrogation de bases de données. L'interprétation du sens d'expressions en langage naturel dans de tels systèmes est, en général, restreinte par le domaine d'application.

Il existe actuellement peu de travaux portant sur la sémantique hors contexte et hors domaine d'application. Notre étude a pour but d'analyser les relations de sens existant entre les mots de groupes nominaux complexes de la forme 'Nom de Nom'[1] de manière à obtenir une couverture maximale de la langue dans un contexte syntaxique restreint. Les différents sens de telles expressions peuvent être mis en évidence par leur traduction dans une autre langue. Par exemple, 'les livres de Sartre' se traduira en Allemand par 'Sartres Bücher' si la relation sémantique est une relation de possession et par 'Die Bücher von Sartre' s'il s'agit d'une relation de production (auteur).

L'application d'un tel analyseur sémantique à la machine à dicter qui fonctionne à base de modèles biclasses ou triclasses essentiellement syntaxiques, permet de sélectionner parmi les expressions reconnues, celles qui sont porteuses de sens et celles qui ne le sont pas.

Dans la première partie, nous allons présenter l'analyseur sémantique et plus particulièrement, ses fondements théoriques, son application spécifique aux expressions à valeur prédicative et son implémentation. Dans le deuxième paragraphe, nous présenterons les fonctionnalités de la machine à dicter et enfin, dans le troisième, les résultats obtenus par l'intégration de l'analyseur sémantique à la machine à dicter.

1. L'ANALYSEUR SEMANTIQUE

1.1. FONDEMENT THEORIQUE

La première étape de notre travail a consisté en l'analyse d'une typologie des expressions de la forme NdeN [Lejosne-91], qui a montré la nécessité de décrire sémantiquement les mots d'une manière très fine. C'est pourquoi nous avons fondé notre étude sur la théorie componentionnelle de F. Rastier [Rastier-87, Rastier-89]. Cette théorie s'appuie sur les notions de sémème et de sème. Un sémème est le contenu sémantique d'un morphème et un sème est caractérisé comme étant

[1] Nous utiliserons indifféremment les notations NdeN, Nom de Nom ou N1 de N2 pour décrire de telles expressions.

l'extrémité d'une relation fonctionnelle binaire entre sémèmes. Ou encore d'après Tutescu [Tutescu-74] cité dans [Rastier-87] : " l'unité minimale de sens, le trait pertinent du contenu sémantique, l'invariant de sens s'appelle marque sémantique, marqueur sémique ou sème...".

On peut distinguer deux classes de sèmes : les sèmes inhérents (relevant du système fonctionnel de la langue) et les sèmes afférents (relevant de normes socialisées voire idiolectales). Les sèmes sont différenciés par leur niveau de généricité, on distingue : les sèmes macrogénériques (ex : /humain/, /animé/), les sèmes mésogénériques (ex : /alimentation/), les sèmes microgénériques (ex : /partie-du-corps/) et enfin les sèmes spécifiques (ex : /fonctionnel/). La description sémantique d'un mot est obtenue par la définition de trois groupes de sèmes :

- le **taxème** contient les sèmes spécifiques et microgénériques. Les sèmes microgénériques servent à regrouper au sein d'un même taxème des éléments voisins alors que les sèmes spécifiques servent à les différencier ;
- le **domaine** est un groupe de taxèmes (caractérisé par un sème mésogénérique) tel que dans un domaine il n'existe pas de polysémie ;
- la **dimension** est une classe de généralité supérieure. Elle inclut des sémèmes comportant un même trait générique (sème macrogénérique). Les dimensions peuvent être articulées entre elles par des relations de disjonction exclusive (ex : /animé/ vs /inanimé/).

EXEMPLE :
'cuiller' est caractérisé par l' ensemble de sèmes génériques suivants (d'après [Rastier-87]):

- taxème : /couvert/
- domaine : /alimentation/
- dimension : /concret/ /inanimé/

Une telle représentation permet d'obtenir une description fine du sens et de définir un ensemble de traits spécifiques à l'étude envisagée.

1.2. APPLICATION AUX PREDICATIFS

Dans un premier temps, nous avons focalisé notre étude sur le traitement des groupements fonctionnels N1deN2 tels que le N1 se comporte de manière prédicative. Les exemples suivants illustrent les trois cas envisagés :

EXEMPLE :

- l'achat de Pierre
  -> action : *Pierre achète quelque chose*
  -> résultat : *ce que Pierre achète*
- le conducteur du camion
  -> agent : *celui qui conduit le camion*

Dans la typologie, il existe trois grandes classes mettant en relation un N1 prédicatif et un N2 argumentatif. Ces trois classes correspondent au cas où N2 est agent (ex : l'achat de Pierre), au cas où N1 est un prédicat et N2 est objet (ex : l'achat de la voiture) et enfin le cas où N1 est un agent et N2 un objet (ex : le conducteur du camion).

Il nous faut déterminer la classe d'appartenance de chaque groupement dont le N1 est de type prédicatif. Pour cela il est nécessaire de connaître la structure sémantique associée au prédicat concerné (i.e. ses arguments) ainsi que les contraintes qui leur sont rattachées. A cette fin, nous avons défini un ensemble de cadres sémantiques, chaque cadre décrivant une classe de prédicats ayant le même comportement sémantique. Comme les seuls arguments que peuvent instancier le N1 et le N2 sont : le prédicat, le nominatif et l'accusatif, les descriptions des cadres sémantiques seront restreintes à ces trois cas (au sens d'une grammaire de cas).

EXEMPLE :

Cadre_achat est un cadre décrivant les arguments des prédicats ayant le même comportement que **acheter**.

Ce cadre possède deux arguments :

- un nominatif dont la dimension doit prendre la valeur : ***humain***
- un accusatif dont la dimension doit prendre la valeur : ***concret et non humain***

Dans cet exemple, dans un souci de simplification, nous ne nous plaçons pas dans le contexte historique où l'on achetait des personne (ex : l'achat de l'esclave) qui relève de la culture mais pas du modèle fonctionnel de la langue[2]. De même, nous n'envisageons pas les cas métaphoriques, tels que ***l'achat du maire par la Mafia***, ou ***l'achat de son silence***.

Les prédicatifs *don, vente, apport...* appartiennent à la classe décrite par le cadre *Cadre_achat*..

En plus du cadre sémantique associé au prédicat, il est indispensable de posséder des informations concernant le comportement de celui-ci dans le contexte d'utilisation que nous étudions. En effet, des prédicats possédant les mêmes arguments nominatif et accusatif, pourront accepter en position de N2, pour l'un, indifféremment le nominatif ou l'accusatif (ex : l'achat de Pierre, l'achat de la voiture), pour un autre n'autoriser que l'accusatif (ex : l'abolition de l'esclavage). Nous avons recensé trois comportements différents :

- **PLEIN** : n'importe quel argument peut être instancié par le N2 ;
- **REDUIT** : l'instanciation est réduite à l'accusatif ;
- **PLEIN_REFL** : le prédicat a à la fois le comportement d'un plein et d'un réfléchi, dans le cas du réfléchi, l'accusatif est égal au nominatif.
  ex : - l'abonnement de Pierre par Marie
  - l'abonnement de Pierre.

Nous pouvons considérer, dans notre application, que les traits /plein/, /réduit/ et /plein_refl/ permettent de regrouper au sein d'un même taxème les mots ayant un même comportement. La gestion des prédicats

[2] Ce qui implique en substance qu'il est encore difficile d'envisager une étude qui aurait une couverture totale de la lengue française actuelle.

ayant un aspect résultatif au sens où nous l'avons défini, passe par la définition d'un trait **spécifique** /résultatif/, en effet, il n'existe pas de classes correspondant aux résultatifs et aux non-résultatifs, mais dans chaque classe que nous avons définie peut se trouver des mots comportant ce trait et d'autres ne l'ayant pas.

Les taxèmes et les dimensions vont nous permettre de donner des définitions sémantiques des mots dans le lexique (les domaines ne sont pas utilisés pour les prédicatifs). Pour certains lexèmes, en particulier pour les agentifs, il nous a paru nécessaire d'introduire deux noyaux sémantiques dans la définition du mot, un noyau principal (NP) spécifique à l'agent et un noyau secondaire (NS) spécifique au prédicat qui lui est associé. La figure 1 nous donne la définition du mot 'conducteur'. Nous remarquerons que le lien avec le cadre sémantique associé au prédicat se fait par l'intermédiaire de son nom.

CONDUCTEUR

Noyau Principal : /conducteur/
taxème : comptable
dimension : +humain
Noyau Secondaire : /conduire/
taxème : plein
dimension : notion
Cadre Associé : cadre_achat =>
contraintes : sème spécifique (ACC) = /véhicule/

Figure 1. description sémantique de 'conducteur'

1.3. IMPLEMENTATION

Les sources de connaissance sont :

- les cadres sémantiques des prédicats qui permettent de définir les contraintes sémantiques portant sur chaque argument, nous en donnons un exemple à la figure 2 ;

cadre_achat
Nominatif ->dimension : +humain
Prédicat ->taxème : prédicat
Accusatif ->dimension : concret et non +humain

Figure 2. Cadre sémantique du prédicat 'achat'

- le lexique qui fournit la description sémantique des mots telle que nous l'avons décrite dans la figure 1 ;
- un ensemble de règles qui oriente l'analyse du N1deN2 en fonction des caractéristiques propres aux deux mots et de valider l'interprétation au moyen d'une fonction d'appariement qui teste si les contraintes sémantiques apportées par la règle et par le cadre sont vérifiées. La figure 3. nous montre un exemple de règle utilisée dans la recherche de la relation sémantique existant entre les deux groupes N1 et N2.

Origine :
cadres_associés au N1
Condition d'application
Taxème (noyau secondaire de N1)
a la valeur /PLEIN/
Actions possibles :
instanciation :
- prédicat = Noyau Secondaire de N1
- nominatif = Noyau Principal de N1
- accusatif = Noyau Principal de N2
schéma associé dans la typologie :
schéma subjectif (N1 sujet)
résultat de l'analyse : Noyau Principal de N1

Figure 3. Exemple de règle d'interprétation

La figure 4. schématise l'analyse du groupement "le conducteur du camion" grâce à la règle et au cadre définis ci-dessus ainsi qu'à la description des mots *conducteur* et *camion*.

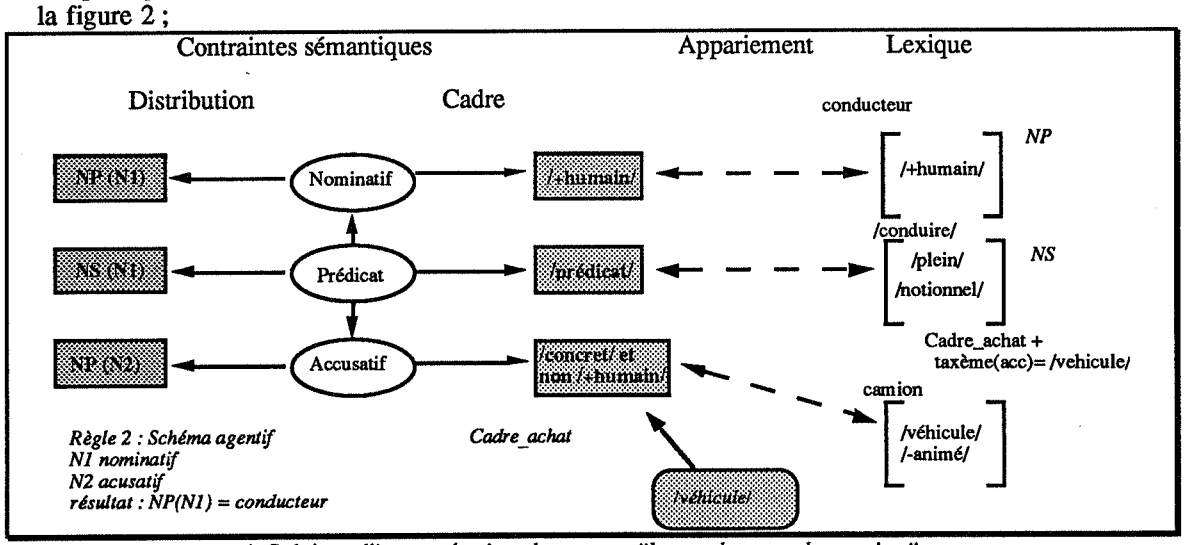


Figure 4. Schéma d'interprétation du groupe "le conducteur du camion"

2. LA MACHINE A DICTER

On confond souvent système de reconnaissance de la parole (SRP) et machine à dicter (MAD). S'il est vrai qu'une machine à dicter est fortement dépendante de son SRP, elle a cependant besoin d'un certain nombre d'outils pour pouvoir fonctionner en tant que

telle. Un simple SRP n'est pas en mesure d'apporter des modifications sur le texte produit par la MAD. D'où l'intérêt d'un éditeur. Cet éditeur peut être vocal ou non.

MAUD (Machine AUtomatique à Dicter) est un prototype de machine à dicter acceptant en entrée le langage naturel (aucune restriction syntaxique n'est imposée). Très souvent l'utilisation du langage naturel s'accompagne par l'utilisation d'un grand vocabulaire. En effet, les systèmes de reconnaissance actuels (et pour un bon nombre d'années à venir) ne peuvent reconnaître un mot si celui-ci n'appartient pas au lexique. La réalisation d'une telle machine a nécessité l'utilisation de quatre composantes : la comosante acoustico-phonétique, la composante lexicale, la composante linguistique, et l'éditeur associé à la MAD [smaïli 91b]. On retrouve ces quatre composantes dans le schéma de l'architecture générale de MAUD de la figure 5.

2.1 LA COMPOSANTE ACOUSTICO-PHONETIQUE DE MAUD

Cette composante est fondée sur le décodeur acoustico-phonétique APHODEX [François 90]. La première tâche de MAUD est de fournir deux treillis phonétiques à partir du treillis phonétique d'APHODEX. Le premier est appelé treillis phonétique d'acceptation (TPA). Ce treillis est construit à partir des nœuds de chaque segment du treillis d'origine, en gardant les étiquettes qui ont un coefficient de vraissemblance supérieur à un certain seuil. Le deuxième est dit treillis phonétique de rejet (TPR). Il est composé des étiquettes phonétiques qui ont été reconnues avec un mauvais score. Ce treillis est utilisé pour rejeter des hypothèses de substitution émises par les niveaux supérieurs.

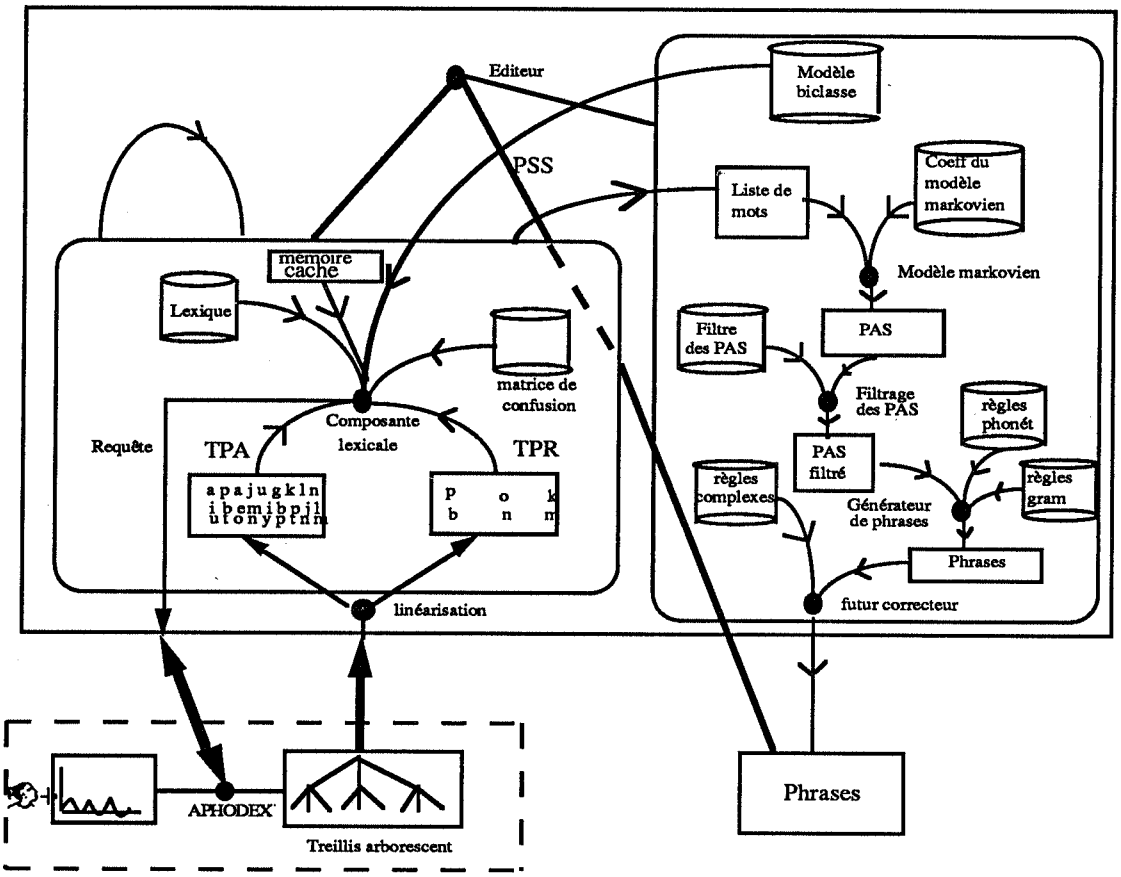


Figure 5 : Architecture générale de MAUD.

Nous détaillons dans ce qui suit les composantes : acoustico-phonétique, lexicale, et syntaxico-sémantique et nous montrerons à quel niveau intervient l'analyseur sémantique 'NdeN' dont il est question dans cet article.

2.2 LA COMPOSANTE LEXICALE

La composante lexicale joue un rôle central dans MAUD puisqu'elle s'articule avec la composante linguistique et la composante acoustico-phonétique avec lesquelles elle interagit pour identifier dans le continuum de parole les mots pouvant être mis en correspondance avec le signal vocal.

Le lexique de MAUD est composé de 37000 entrées lexicales. Il ne faut pas considérer ce lexique comme une simple liste de mots, mais comme une base de données de laquelle on peut extraire un grand nombre d'informations concernant chacune de ces entrées. La tâche d'identification ne peut s'effectuer sans une organisation efficace du lexique. Les fonctions d'accès doivent en faciliter la mise en œuvre de filtres dont la combinaison doit permettre d'extraire du lexique des sous-ensembles aussi restreints que possible. Les informations lexicales du lexique de MAUD sont réparties en deux catégories permettant à la composante lexicale de MAUD d'agir sur deux niveaux : infra-lexical, et supra-lexical. La composante supra-lexicale permet les interactions avec les niveaux syntaxico-sémantiques, alors que la composante infra-lexicale facilite la communication avec le niveau acoustico-phonétique.

La mise en place de procédures d'accès au lexique doit être aussi efficace que possible tant les accès sont nombreux pendant la reconnaissance. Pour ce faire, un certain nombre de filtres très complexes sont mis en œuvre dans MAUD permettant une efficacité, à la fois sur la rapidité et sur la capacité à extraire des sous-vocabulaires restreints [Smaïli 92].

2.3 LA COMPOSANTE SYNTAXICO-SEMANTIQUE

Le rôle de la syntaxe en reconnaissance de la parole est de participer au choix du prochain mot à reconnaître et à l'élimination d'un certain nombre d'hypothèses. La constitution de phrases d'une langue n'est pas une simple combinaison de mots, pris dans n'importe quel ordre, mais un mécanisme de construction de phrases très précis. En traitement de langue naturelle, on ne sait toujours pas fournir un modèle linguistique permettant de traiter automatiquement la langue. C'est pour cette raison, que les informaticiens partent du principe que la probabilité de production d'un mot dépend conditionnellement de toute la première partie de la phrase, pour proposer un modèle permettant de traiter la langue. D'après cette constatation, il est naturel de penser à l'utilisation d'un modèle probabiliste. En effet, la composante syntaxico-sémantique de MAUD est composée d'un modèle markovien comprenant 6000 états et 37000 transitions. Ce modèle est augmenté d'un certain nombre de règles grammaticales et phonologiques permettant de prendre en compte les phénomènes linguistques qui ne peuvent l'être par le modèle probabiliste. MAUD est composée de sept modules syntaxico-sémantiques : le préprocesseur syntaxico-sémantique, le processeur stochastique, le filtre des patrons syntaxiques [Smaïli 91a], le générateur de phrases, le filtre grammatical, et le filtre phonologique.

Ces sept modules agissent de manière pyramidale. Autrement dit, lorsque les solutions proposées à un certain niveau arrivent au niveau supérieur, elles sont filtrées (donc réduites) et envoyées de nouveau au niveau immédiatement supérieur. Ce filtrage multi-niveaux assure une bonne réduction de l'espace de solutions. Cependant, et ce malgré l'existence de ces sept modules syntaxico-sémantiques, à cause de l'imperfection du décodage acoustico-phonétique et du non recouvrement total de la langue du modèle probabiliste, les solutions proposées sont en nombre important. Pour réduire le nombre de propositions, une première solution consiste à introduire un analyseur sémantique permettant de traiter les groupes 'NdeN' prédicatifs. L'apport d'un tel analyseur est très important comme le montre les résultats des prochains paragraphes.

3. INTEGRATION DE L'ANALYSEUR SEMANTIQUE A LA MACHINE A DICTER

Nous avons utilisé l'analyseur sémantique des expressions en NdeN comme un filtre agissant sur les résultats obtenus par la machine à dicter. Nous avons analysé quatre phrases contenant les expressions suivantes : 'l'achat de Chantal', 'la chute de l'enfant', 'le chauffeur de taxi' et 'l'achat du livre'. Nous avons extrait des résultats fournis ceux ayant une valeur prédicative et nous les avons soumis à l'analyseur sémantique. Sur 51 expressions proposées, 20 possèdent une valeur prédicative parmi celles-ci 9 ont été validées et correspondent aux expressions qui intuitivement paraissent correctes. L'inconvénient principal de la restriction aux expressions à valeur prédicative, est que l'on ne peut pas déterminer si une expression n'a pas été validée parce qu'elle n'est pas porteuse de sens ou parce qu'elle n'est pas prédicative. L'intérêt dans ce cas de figure est d'utiliser la validation sémantique comme un facteur intervenant sur le score de reconnaissance, ce qui augmente ainsi la convivialité de l'interface avec l'utilisateur qui aura l'avantage de trouver les meilleures expressions en tête de liste.

Ceci n'étant malgré tout pas entièrement satisfaisant, nous avons décidé d'étendre l'analyse à toutes les expressions en NdeN. De la même manière que nous avons associé un cadre à un prédicat, nous avons défini un ensemble de cadres associés aux différentes classes de la typologie (ex : appartenance, propriété ...). Certaines relations n'ont cependant pas pu être explicitées car elles ne dépendent pas de la sémantique mais de connaissances encyclopédiques pour lesquelles il faudra envisager un traitement particulier (ex : relation de production : les livres de Sartre, relations de partie-tout : le pied de la table). Certains cadres sont quant à eux associés directement aux mots même si ceux-ci ne sont pas des prédicatifs (ex : spécialiste -> cadre connaitre). Après l'intégration de ces cadres, nous avons pu constater que la typologie était incomplète d'où le rejet de certaines expressions pourtant valides. Sur les 51 expressions de départ, seules 16 d'entre elles ont été validées, et toutes sont porteuses de sens. Parmi les expressions rejetées, il y en a 6 qui font partie des expressions qui peuvent entrer dans la catégorie partie-tout et qui n'ont donc pas été traitées, la seule solution envisagée actuellement est de ne pas les rejeter même si l'on est incapable de

déterminer si elles ont un sens ou non. Sur les 51 expressions de départ, nous nous retrouvons donc avec 22 expressions conservées et 29 rejetées, ce qui fait un gain de plus de 50%. Parmi les expression validées, nous trouvons par exemple : *l'achat du fauteuil, l'achat du sapin* (pour la phrase contenant *l'achat de Chantal)*, ou encore *la chute de l'avocat, la chute de l'assassin* (pour *la chute de l'enfant*), et parmi celles rejetées nous avons: *l'agent de chagrin, l'enfant de sapin, le chauffeur de pêcher...*

CONCLUSION

L'étude qui a été faite montre l'utilité d'une analyse sémantique comme filtre des résultats de la machine à dicter. Les jeux d'essais effectués ne permettent pas d'établir le gain moyen obtenu par cette analyse sur un grand corpus mais les résultats obtenus sont tout de même pertinents et montrent la validité des traitements. L'application à la machine à dicter a permis de mettre en évidence un certain nombre de classes ne figurant pas dans la typologie et qu'il faudra prendre en compte dans la suite des travaux. Une extension de l'étude faite sur les expressions prédicatives consistant en la définition des modules casuels complets associés aux prédicats (et non plus limités au nominatif et à l'accusatif) permettra d'élargir l'analyse sémantique à des expressions englobant des NdeN et de résoudre ainsi un certain nombre d'ambiguïtés d'interprétation possibles (la conduite de Paul est sportive, la conduite de Paul à la gare m'a pris deux heures).

Dans la version actuelle, l'analyseur sémantique sert uniquement à valider ou invalider des expressions reconnues par la machine à dicter. il serait intéressant maintenant de l'intégrer à celle-ci de manière à pouvoir faire des prédictions lors de la reconnaissance de phrases et ainsi, d'une part, ne construire que des groupes corrects, et d'autre part, de gagner du temps par la réduction par contraintes sémantiques du lexique à consulter.

BIBLIOGRAPHIE

[François-90] D.François, D.Fohr " Première évaluation d'APHODEX, système expert pour le décodage acoustico-phonétique de la parole continue", XVIII Journées d'Etudes sur la parole, 1990.

[Lejosne-91] Lejosne J.-C., Klein J., Lauvray J., Romary L., "Typologie des groupes nominaux complexes", Colloque *Lexique et Inférence*, Metz, 1991.

[Rastier-87] Rastier F. *Sémantique interprétative*, puf , Formes sémiotiques(1987).

[Rastier-89] Rastier F. *Sens et Textualité*, Hachette, Paris (1989)

[Smaïli-91a] K.Smaïli, F.Charpillet, JM.Pierrel, JP.Haton " A continuous speech recognition approach for the design of a dictation machine", European Conference on SSpeech Technology, pp953-956, Genova 1991.

[Smaïli-91b] K.Smaïli " Conception et réalisation d'une machine à dicter à entrée vocale destinée aux grands vocabulaires: Le système MAUD", Thèse de Doctorat de l'université de Nancy I, 1991.

[Smaïli-92] K.Smaïli, F.Charpillet, JM.Pierrel, JP.Haton " La composante lexicale de la machine à dicter MAUD", Séminaire lexique. Communication Homme-Machine Pôle langage naturel, pp46-57, 1992

19es J.E.P. - Bruxelles, 19 au 22 mai 1992.

VINICS : UN SYSTEME ADAPTATIF DE RECONNAISSANCE DE LA PAROLE CONTINUE

GONG Yifan et MOURIA Fériel

C.R.I.N/C.N.R.S & I.N.R.I.A-Lorraine
BP 239, 54506 Vandœuvre, France.

Résumé

On se propose dans cet article de présenter un système de reconnaissance de la parole continue fondé sur une nouvelle approche. Cette approche consiste à utiliser une fonction représentant les plausibilités d'observer chaque symbole phonétique, et à trouver la phrase qui maximise la somme des plausibilités de chaque symbole constituant la parole présentée en entrée.

Le système a un taux de reconnaissance de 95% au niveau mono-locuteur et un taux de 95.2% en mode adaptatif (il s'agit d'adapter le système à de nouveaux locuteurs). Les résultats des expériences montrent que cette méthode donne une meilleure précision au niveau de la reconnaissance et nécessite 1/20 du temps de calcul des méthodes traditionnelles fondées sur la programmation dynamique.

1 INTRODUCTION

Cet article décrit l'approche que nous avons adopté pour résoudre le problème de la reconnaissance de la parole continue et présente le système : VINICS[1] fondé sur cette approche. Nous décrivons également l'adaptation du système à de nouveaux locuteurs et les résultats obtenus.

Le but du système est de reconnaître la parole continue en optimisant la plausibilité cumulée des symboles composant la phrase prononcée.

VINICS est un système guidé par la syntaxe, pour un vocabulaire moyen et adaptable à de nouveaux locuteurs.

[1] VINICS : Voice Interface based on the Notion of Image Center of Symbols

Les connaissances sémantiques, syntaxiques, lexicales, phonologiques et phonétiques sont compilées sous forme d'un automate non-déterministe. Il est destiné à différentes applications et est fondé sur une grammaire à contexte-libre.

Actuellement, le système utilise une grammaire sémantique pour la langue française comportant 1500 règles avec un vocabulaire de 1200 mots. Cette grammaire peut engendrer plus de 10^{13} phrases différentes concernant les réservations des places dans les trains, le contrôle d'un robot, les demandes de renseignements dans une bibliothèque, les noms de villes françaises, les noms de pays, les nombres de 0 à 9999 et les numéros de téléphone [7].

Nous allons introduire dans le paragraphe suivant la formulation mathématique du principe du système. le troisième paragraphe sera consacré à la description détaillée de la résolution de cette formulation. Puis nous présenterons l'adaptation du système à de nouveaux locuteurs. Enfin nous concluerons par les expériences d'évaluation du système, au niveau mono-locuteur et en mode adaptatif.

2 FORMULATION MATHEMATIQUE

La reconnaissance de la parole consiste essentiellement à projeter un signal de parole sur une séquence de symboles primitifs comme les symboles phonétiques. En parole continue, les images

acoustiques de quelques symboles consécutifs ne sont pas concaténées entre elles, mais étendues le long de l'axe du temps. Quand on compare ces images à celles des élocutions de phonèmes isolés, on remarque qu'elles sont très influencées par le contexte (les symboles voisins). En revanche, le centre d'une image est généralement moins influencé par ce contexte.

Afin de contourner ce problème de l'influence contextuelle sur les limites de l'image acoustique d'un phonème, nous adoptons une technique qui met l'accent sur le centre de l'image acoustique des symboles phonétiques en parole continue. Cette technique calcule à chaque prélèvement du signal acoustique la plausibilité de chaque phonème sans spécifier les instants de début et de fin de ce phonème [6].

Pour réaliser la reconnaissance de la parole continue fondée sur la mesure des plausibilités, nous utilisons une procédure d'alignement temporel pour trouver la phrase la plus plausible. Cette procédure compare, par une recherche exhaustive, les phrases de référence avec la plausibilité des symboles en entrée [1, 2, 8].

Le système utilise la programmation dynamique pour :

- trouver le meilleur chemin liant les pics de deux symboles successifs à partir des fonctions de plausibilité
- trouver le meilleur temps de transition du numéro du prélèvement pour deux pics donnés.

Supposons qu'une phrase ph soit composée d'une séquence de $L(ph)$ symboles phonétiques s_i :

$$ph = s_1, s_2, ..., s_{L(ph)} \qquad (1)$$

On pose, la fonction de vraisemblance suivante : $\mu_{s,n} : s \times n \rightarrow [0,1]$ qui représente la plausibilité du symbole phonétique s à l'instant n.

La reconnaissance de la parole continue consiste à trouver la phrase ph qui maximise le cumul des plausibilités des symboles s_i composant la phrase prononcée :

$$Q(ph) = \max_{\mathcal{Z}(s_k)} \sum_{k=1}^{L(ph)} \sum_{n \in \mathcal{Z}(s_k)} \mu_{s_k,n} \qquad (2)$$

$\mathcal{Z}(s_k)$ est la région de l'image acoustique du symbole s_k où $\mu_{s,n}$ a la plus grande valeur.

Le problème de la reconnaissance de la parole continue se résume donc à calculer la fonction $\mu_{s,n}$ et à trouver la phrase $\hat{ph}$ de l'ensemble de toutes les phrases possibles PH qui donne la meilleure plausibilité cumulée, c'est à dire :

$$\hat{ph} = argmax_{ph \in PH} Q(ph) \qquad (3)$$

3 SOLUTION PROPOSEE

3.1 La recherche guidée par les pics

Pour un symbole spécifique s, la plausibilité $\mu_{s,n}$ est calculée à l'instant n, donc la fonction de plausibilité varie avec le temps. L'instant de temps d'un maximum local de la fonction $\mu_{s,n}$ est appelé un pic de s.

Nous appelons "indice" le numéro du prélèvement dans la fonction de vraisemblance $\mu_{s,n}$. Soit $p_{s,i}$ l'indice du i^e pic du symbole s. Etant donné que dans la région $\mathcal{Z}(s)$ le symbole s a relativement grande plausibilité, on peut considérer, qu'à l'intérieur de $\mathcal{Z}(s)$, il existe au moins un pic de s. Le chemin de recherche de l'ensemble des régions $\mathcal{Z}(s)$ qui donne la meilleure plausibilité cumulée peut être guidé par les pics des symboles successifs.

Soit m_s le nombre total des pics d'un symbole s dans une phrase prononcée. En fonction des pics, la procédure de recherche de la phrase la plus plausible $\hat{ph}$ peut être décomposée en deux processus successifs de maximisation [4] :

- choisir dans des symboles successifs s_j, un ensemble optimal de pics noté $\mathcal{P}$:

  $$\mathcal{P} = \{P_{s_j} \in \{p_{s_j,1}, p_{s_j,2}, ... p_{s_j,m_{s_j}}\}|_{j \in [1,L(ph)]}\}$$

  avec :

  $$P_{s_1} < P_{s_2} ... < P_{s_{L(ph)}}.$$

- déterminer un ensemble optimal de chemins d'indice de transition à partir du symbole s_k jusqu'au symbole s_{k+1} noté $\mathcal{X}$:

  $$\mathcal{X} = \{x_k|_{k \in [1,L(ph)]}\}$$

  avec :

  $$P_{s_k} \leq x_k < P_{s_{k+1}}.$$

Dans la suite, nous décrivons brièvement ces deux processus, pour plus de détails le lecteur peut se référer à [4].

3.2 Trouver la meilleure séquence de pics

Le premier processus de maximisation consiste à trouver le meilleur chemin qui relie les pics des fonctions de plausibilité des symboles successifs.

En utilisant la programmation dynamique, nous introduisons la fonction récursive $S(s_j, i)$:

$$S(s_j, i) = \max_{p_{s_{j-1},k} < p_{s_j,i}} [S(s_{j-1}, k) + Optsum(s_{j-1}, s_j, p_{s_{j-1},k}, p_{s_j,i})] \quad (4)$$

$$1 \leq j \leq L(ph), \ 1 \leq i \leq m_{s_j}.$$

$S(s_{L(ph)}, m_{s_{L(ph)}})$ donne le résultat du processus d'optimisation. La fonction *Optsum* donne la somme maximale de plausibilité quand le chemin va de s_{j-1} à s_j dans l'intervalle de temps $[p_{s_{j-1},k}, p_{s_j,i}]$. Cette fonction est détaillée dans le paragraphe qui suit.

La figure 1 montre le diagramme de transition d'état de ce processus.

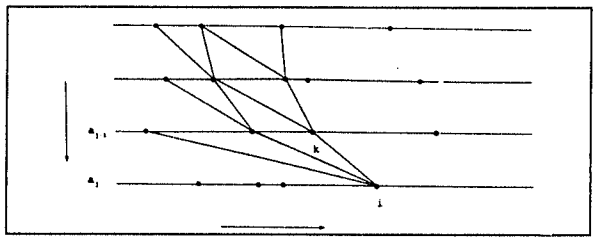


Figure 1: Le diagramme de transition d'état pour la recherche du meilleur chemin reliant les pics des symboles successifs. La programmation dynamique est utilisée pour calculer $S(s_j, i)$ dans l'équation 4. Les points représentent les positions des pics dans le temps.

3.3 Trouver l'indice de la meilleure transition

Dans le paragraphe précédent, nous avons introduit la fonction *Optsum* qui maximise la plausibilité cumulée de deux symboles successifs s_p (précédent) et s_c (courant) en ajustant le chemin de transition de s_p à s_c dans l'intervalle $[a, b]$:

$$Optsum(s_p, s_c, a, b) = \max_{a \leq x < b} (\sum_{n=a}^{x-1} \mu_{s_p,n} + \sum_{n=x}^{b-1} \mu_{s_c,n}) \quad (5)$$

a et b sont respectivement les indices des pics s_p et s_c. Cette optimisation revient à retrouver un indice $n_{opt}(s_p, s_c, a, b)$ tel que le chemin passant par des symboles successifs en allant de s_p à s_c fournit le maximum de la partie droite de l'équation 5 :

$$n_{opt}(s_p, s_c, a, b) = argmax_{a \leq x < b} (\sum_{n=a}^{x-1} \mu_{s_p,n} + \sum_{n=x}^{b-1} \mu_{s_c,n}) \quad (6)$$

La partie droite de l'équation 5 peut donc être calculée en un seul passage et s'écrit de la manière suivante :

$$\sum_{n=a}^{b-1} \mu_{s_c,n} + \max_{a \leq x < b} \sum_{n=a}^{x-1} (\mu_{s_p,n} - \mu_{s_c,n}) \quad (7)$$

La figure 2 illustre cette recherche.

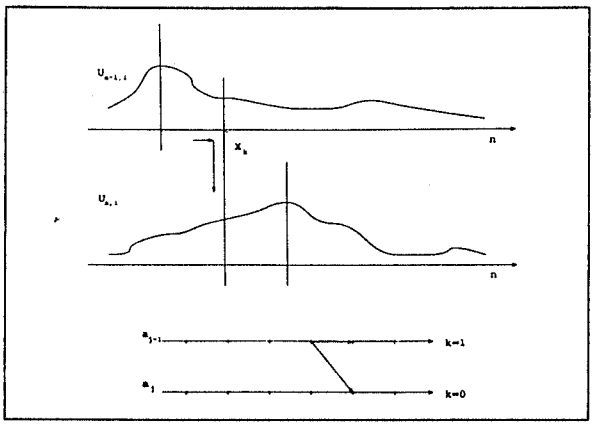


Figure 2: Le diagramme de transition d'état pour la recherche de l'indice de temps qui maximise la plausibilité cumulée entre deux pics.

4 ADAPTATION

Pour s'adapter à un nouveau locuteur, le système utilise un certain nombre de phrases prononcées par ce dernier.

VINICS dispose de la prononciation des phrases par un locuteur de réfèrence pour étiqueter les phrases du nouveau locuteur. Le principe de l'étiquetage consiste à :

- effectuer un alignement dynamique des phrases du nouveau locuteur avec celles du locuteur de référence. Cet alignement est réalisé dans un espace paramétré indépendamment du locuteur,
- extraire automatiquement les références des phonèmes en effectuant une projection des segments étiquetés pour le locuteur de référence sur les phrases du nouveau locuteur.

Pour mesurer les distances dans la procédure d'alignement temporel, nous avons besoin des vecteurs E_n, E_f et de leurs dérivées par rapport au temps. E_n représente l'énergie moyenne entre 200 Hz et 3200 Hz et E_f est donnée par le rapport $\frac{e_2}{e_1+e_2}$ où e_1 représentant l'énergie moyenne comprise entre 150 Hz et 1100 Hz et e_2 celle comprise entre 3000 Hz et 6000 Hz.

La figure 3 montre la valeur de E_n en fonction du temps ainsi que le résultat de l'alignement automatique pour les locuteurs *jph* (comme référence) et *jmp* (comme test).

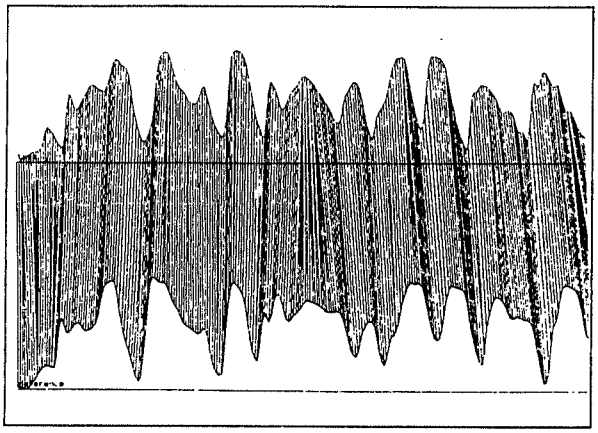


Figure 3: Les contours du vecteur E_n et le résultat de l'alignement temporel automatique des prononciations de la phrase " Guy a péri bêtement du diabète en Italie " par les locuteurs *jph* et *jmp*.

La figure 4 présente les spectrogrammes et les projections d'étiquettes pour une phrase prononcée par ces deux locuteurs.

Quelques erreurs d'étiquetage sont possibles car, durant la phase d'alignement, les variations phonologiques ne sont pas prises en compte.

5 RESULTATS

Dans ce paragraphe nous présentons les résultats des tests de reconnaissance au niveau monolocuteur et ceux concernant l'adaptation du système à de nouveaux locuteurs.

VINICS a été testé en utilisant 10 locuteurs masculins, chacun ayant prononcé, 3 à 5 fois, un corpus composé de 17 phrases. La première répétition est utilisée pour l'apprentissage, les autres ont servi de données de test. La reconnaissance au niveau des mots pour les 10 locuteurs est donnée dans le tableau 1 .

Les taux d'erreurs sont compris entre 1.3% (*jph*) et 9.1 (*jlc*)%, avec un taux moyen inférieur à 5%.

Pour adapter le système à de nouveaux locuteurs, on a été amené à faire de la reconnaissance du locuteur. Dans un premier temps, nous avons réalisé un module de reconnaissance du locuteur indépendamment du texte prononcé. Dans ces conditions, le temps mis pour reconnaître un nouveau locuteur a été relativement long mais le taux de reconnaissance de la parole atteint 99.5%. Pour plus de détails sur ces expériences, le lecteur peut se référer à [3].

Ensuite, nous sommes interessés au cas où la reconnaissance du locuteur dépend du texte prononcé. Pour cela, nous nous sommes imposés d'utiliser un texte d'apprentissage très court [5]. Il est demandé au nouveau locuteur de prononcer au maximum 2 mots et le module d'adaptation cherche parmi les locuteurs de référence celui qui a les caractéristiques phonétiques les plus proches. Le taux de reconnaissance a été de 91%. Nous remarquons une légère baisse du taux de reconnaissance mais l'avantage, par rapport au premier module, a été de gagner du temps au niveau de la reconnaissance du locuteur.

Des expériences ont été réalisées en mode adaptatif avec 5 locuteurs masculins, les résultats obtenus sont regroupés dans le tableau 2 . Pour les 5 locuteurs le taux de reconnaissance est compris entre 91.2% et 100% avec un taux moyen de 95.2%.

6 CONCLUSION

Nous avons présenté dans cet article une nouvelle méthode pour la reconnaissance de la parole continue. Etant donné qu'en parole continue le centre des images acoustiques est moins déformé par le contexte que les extrémités, la reconnaissance basée sur les régions de forte plausibilité est plus efficace que les méthodes traditionelles basées sur les frontières.

Notre principal souci a été de mettre en œuvre et de tester cette nouvelle méthode de reconnaissance de la parole continue. Les premières expérimentations sont très encourageantes, puisque les résultats obtenus indiquent un taux de reconnaissance moyen en mode adaptatif de l'ordre de 95%.

Par rapport à une méthode markovienne, le système demande un nombre de phrases d'apprentissage nettement inférieur (de l'ordre du dixième), ce qui constitue un avantage important pour de grands vocabulaires et pour l'adaptation à de nouveaux locuteurs.

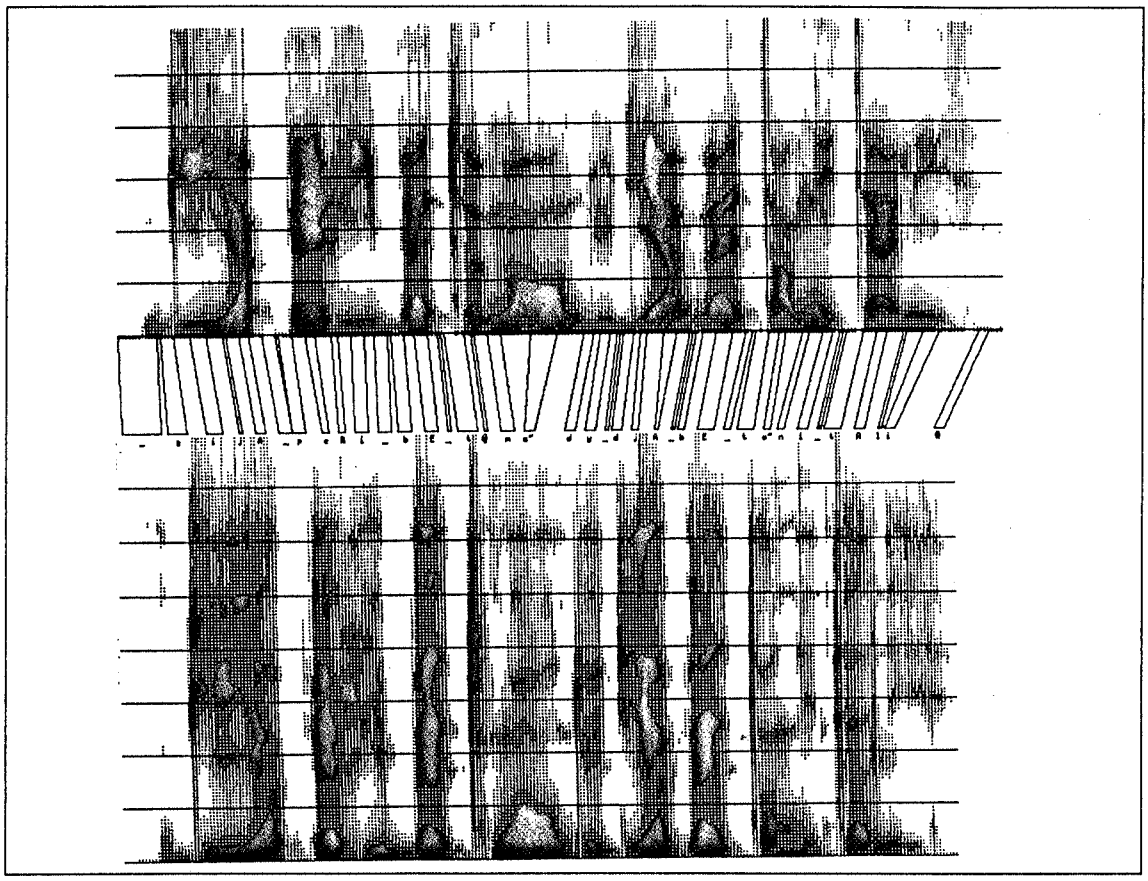


Figure 4: Les spectrogrammes et les résultats de l'alignement temporel automatique des prononciations des deux locuteurs *jph* et *jmp* en utilisant la même phrase de test que dans la figure 3.

locuteur	aq	bz	df	gm	jfm	jg	jlc	jmp	jph	ms
nb. de répititions	4	3	2	3	2	3	3	3	3	2
mots erronés	13	13	12	18	6	23	34	16	5	11
taux d'erreur	2.6%	3.5%	4.8%	4.8%	2.4%	6.2%	9.1%	4.3%	1.3%	4.4%

Table 1: Taux d'erreur de la reconnaissance pour les 10 locuteurs.

locuteur	jph	ms	aq	bz	jfm	moyenne
nb. d'énoncés de test	51	34	68	51	34	47.6
nb. d'énoncés reconnus	51	31	67	47	32	45.6
emps de reconnaissance (s)	21.6	19.6	26.9	27.8	15.6	22.3
taux de reconnaissance (%)	100	91.2	98.5	92.2	94.1	95.2

Table 2: Les résultats d'évaluation du système VINICS en mode adaptatif pour 5 locuteurs masculins.

Références

[1] G. D. FornyJr. The Viterbi algorithm. In *Proc. IEEE*, volume 61, pages 268–278, Mar 1973.

[2] M. Franzini, K. F. Lee, and A. Waibel. Connectionist Viterbi Training: A New Hybrid Method for Continuous speech Recognition. In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing 1990*, volume 1, pages 425–428, Albuquerque, New Mexico, USA, April 1990.

[3] Y. Gong and J.-P. Haton. Text-Independent Speaker Recognition by Trajectory Space Comparison. In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing 1990*, Albuquerque, New Mexico, U.S.A., April 1990.

[4] Y. Gong and J.-P. Haton. Signal-to-string conversion based on high likelihood regions using embedded dynamic programming. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 13(3):297–302, March 1991.

[5] Y. Gong and J.-P. Haton. Non-linear vectorial interpolation for speaker recognition. In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing 1992*, San Francisco, USA, March 1992.

[6] Y. Gong, J.-P. Haton, and F. Mouria. Continuous speech recognition based on high plausibility regions. In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing 1991*, volume 1, pages 725–728, Toronto, Canada, May 1991.

[7] Y. Gong, F. Mouria, and J.-P. Haton. Un système de reconnaissance de la parole continue sans segmentation. In *Actes du $7^{ème}$ congrès Reconnaissance des Formes et Intelligence Artificielle*, volume 3, pages 1191–1203, Paris, France, Nov. 1989. AFCET, INRIA.

[8] N. Morgan and H. Bourlard. Continuous speech recognition using multilayer perceptrons with hidden markov models. In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing 1990*, volume 1, pages 413–426, Albuquerque, New Mexico, USA, April 1990.

19es J.E.P. - Bruxelles, 19 au 22 mai 1992.

RECONNAISSANCE DE LA PAROLE DANS LE PROJET MULTIWORKS

J. CAELEN, E. REYNIER, Ph. VERDIER, A. LICHENE

INSTITUT DE LA COMMUNICATION PARLEE, U.R.A.-CNRS 368
I.N.P.G & UNIVERSITE STENDHAL, 46 Av. FELIX VIALLET
38031 GRENOBLE CEDEX - FRANCE

Résumé

Cet article décrit le système de reconnaissance automatique de la parole mis au point dans le cadre du projet Multiworks (projet ESPRIT II n°2105). Ce système est implanté sur une carte dédiée devant équiper la station de travail multimédia "Multiworks". Il se compose de trois étages de traitement fonctionnellement distincts :

1- l'analyse paramétrique du signal et la reconnaissance proprement dite, effectuée par l'algorithme de Viterbi-Ney (mots enchaînés) dans une perspective ascendante,

2- l'analyse linguistique opérée par un ATN (Augmented Transition Network) —résultant de la compilation d'une grammairre lexicale fonctionnelle— qui gère et contrôle l'émission des hypothèses lexicales faite par l'étage précédent,

3- le filtrag des solutions finales par un algorithme DTW (Dynamic Time Warping) à partir de modèles phonétiiqes, dans une perspective descendante.

1. INTRODUCTION

Le projet Multiworks (ESPRIT II n°2105) faisant suite au projet COCOS (ESPRIT I n°990), a débuté en 1988 avec pour objectif la réalisation d'une station de travail multimédia sur la base des grands standards actuels (Unix, X, etc.). A cette fin, des entrées-sorties vocales (avec reconnaissance et synthèse) s'imposaient : une réalisarion matérielle (carte dédiée) et logicielle a été confiée à un consortium composé de BULL-DEA, CRIN, ICP, IRCAM et OROS. l'ICP a eu en charge la réalisation de la carte (avec OROS) et la mise au point des logiciels de reconnaissance et de synthèse. C'est ce travail que nous décrivons ci-après à l'exclusion de la synthèse (voir pour cela [Bailly, 90]).

2. LA CARTE VOCALE

Le synoptique de la fig. 1 présente le schéma général de la carte vocale VAPS (Voice Audio Processing extension System) qui a été configurée pour permettre :

- une qualité hifi stéréo en entrée-sortie,
- la reconnaissance et la synthèse de la parole en temps réel, le traitement du signal pour des applications musicales.

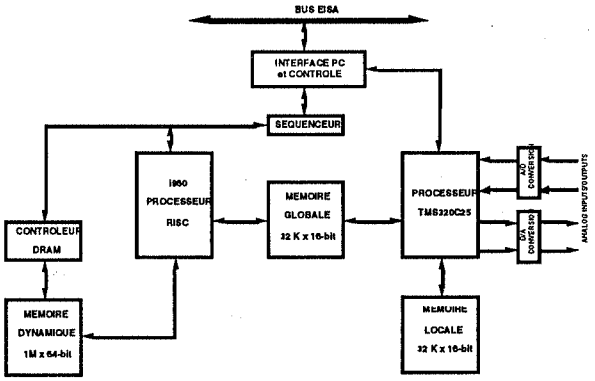


Fig. 1 : Synoptique du système VAPS.

Le système VAPS se compose de deux cartes au format PC connectées sur le bus EISA dans l'espace d'entrées-sorties de la station de travail. Il se compose de différents modules :

• sur la première carte,

- une interface avec la station,
- un processeur général RISC tournant à 32 MHz (INTEL i860) capable d'effectuer les fonctions de traitement de signal de haut niveau, associé à une mémoire dynamique (1 M x 64 bits) directement accessible de la station,
- une interface entre les deux cartes,

• sur la deuxième,

- un double canal analogique-digital pour les entrées-sorties de signal en stéréo. Les deux canaux sont entiérement indépendants du point de vue du contrôle

mais les échantillons restent toujours synchronisés. Chaque canal contient de l'entrée vers la sortie,
- un interrupteur programmable qui assure la connexion de
de l'entrée sur une tension de référence,
- un système de protection contre les variationss de tension,
- un amplificateur (d'impédance d'entrée = 1M Ω)
- deux amplificateurs programmables de 0 à 40 dB (l'un à pas de variation large, l'autre à pas fin),
- un système de correction d'offset,
- deux convertisseurs 16-bit Delta-Sigma (A/D et D/A),
- un amplificateur de sortie à gain fixe (d'impédance = 600 Ω)
- un DSP (Digital Signal Processor) à virgule fixe (Texas Instruments TMS320C25) dédié au contrôle des entrées-sorties mais capable d'effectuer quelques opérations simples de prétraitement comme le filtrage, la pré-emphase, etc. Ce processeur adresse une mémoire locale de 32 K x 16 bit et une mémoire globale de 32 K x 16 bit partagée avec le processeur i860,
- une interface MIDI asynchrone à 31.25 kHz.

Plusieurs prototypes du système VAPS ont été réalisés et les tests ont été effectués avec succès.

3. ARCHITECTURE GENERALE DU SYSTEME DE RECONNAISSANCE

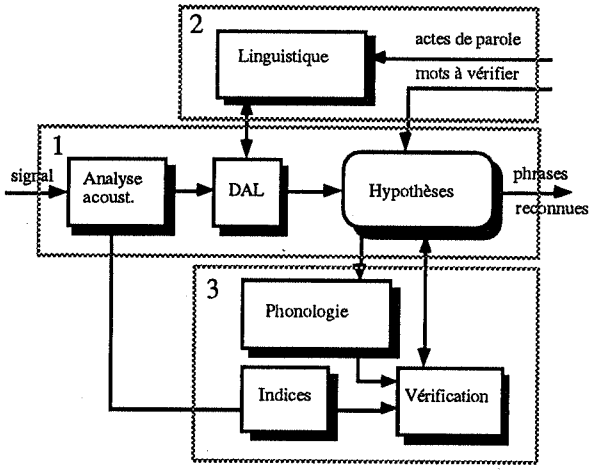


Fig. 2 : Schéma général du système de reconnaissance (DAL = Décodeur Acoustico-Lexical).

Il se compose de trois étages de traitement fonctionnellement distincts :
1- l'analyse paramétrique du signal et la reconnaissance proprement dite, effectuée par l'algorithme de Viterbi-Ney (mots enchaînés) dans une perspective ascendante,
2- l'analyse linguistique opérée par un ATN (Augmented Transition Network) —résultant de la compilation d'une grammairre lexicale fonctionnelle— qui gère l'émission des hypothèses lexicales faite par l'étage précédent. On notera que l'analyseur linguistique peut dynamiquement tenir compte de contraintes provenant de l'extérieur (du module de dialogue notamment). Il peut également dévier dans le tableau des hypothèses des séquences de mots à vérifier,
3- la vérification et le filtrage des solutions par un algorithme DTW (Dynamic Time Warping) à partir de références phonétiques, dans une perspective descendante.
Ces trois modules sont indépendants ce qui permet de configurer à loisir le système de reconnaissance en assemblant les parties (1) ou (1+2) ou (1+2+3) selon la puissance de la machine (ou de la carte) cible et les coûts de développement des bases de connaissance que l'on est prêt à consentir pour réaliser une application particulière nécessitant un vocabulaire et une syntaxe déterminés.

4. LE DECODAGE LEXICAL ASCENDANT

Le décodage lexical ascendant (DAL) [Yé, 90] reprend la majorité des aménagements et des améliorations des chaînes de Markov proposées dans la littérature. Il permet l'apprentissage et la reconnaissance de mots enchaînés prononcés de façon naturelle par plusieurs locuteurs. L'apprentissage se fait sur des mots isolés mais grâce à l'algorithme de Ney modifié et adapté aux chaînes de Markov, la reconnaissance se fait sur des mots connectés en une seule passe, de gauche à droite.

Les caractéristiques de ce système sont les suivantes :
- l'analyse spectrale est réalisée sur des fenêtres temporelles de 128 échantillons de signal numérisé avec possibilité de 3 analyses différentes : LPC (Linear Predictive Coding), cepstre ou modèle d'oreille,
- la quantification vectorielle (QV) utilise la méthode "SPLIT" sur un dictionnaire multi-paramétrique de 128 ou 256 symboles — les vecteurs sont issus de l'analyse spectrale, des informations de durée, de l'énergie et de sa dérivée,
- la reconnaissance est effectuée par l'algorithme de Viterbi-Ney (avec des modèles de Markov à 5 états), le cheminement étant contraint par une matrice de durée [Rabiner, 86], et les modèles de mots étant concaténés pour la reconnaissance de mots enchaînés sous contrôle du niveau linguistique. Les probabiltés de transition d'un mot aux suivants permettent de calculer une probabilité globale pour une séquence de mots, en les combinant aux probabilités de chacun des mots composant la séquence.

Les modifications apportées aux modèles de Markov classiques pour les adapter au cas particulier de la parole continue multilocuteurs sont, outre le nombre d'états et leur topologie linéaire, l'apprentissage multi-observations (utilisation de plusieurs prononciations du même mot pour l'apprentissage d'un seul modèle), un aménagement de calcul pour éviter les dépassements de capacité lors de l'estimation des probabilités, et

l'utilisation de contraintes sur les durées des états. Les contraintes sur les durées ont été introduites dans les HMM [Levinson 86] car il se peut que les probabilités

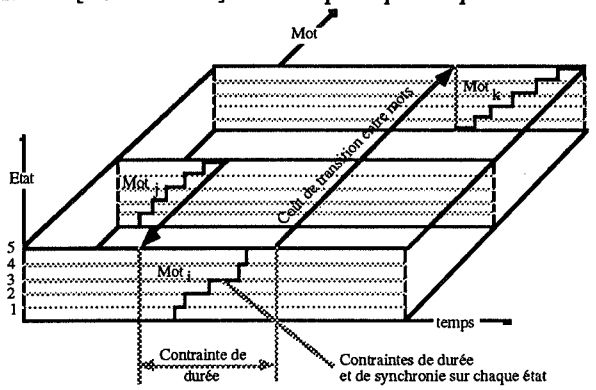


Fig. 3 : représentation d'un "chemin" optimal pour une phrase de trois mots.

d'émission d'un état d'une chaîne et la probabilité de rester dans cet état soient telles que ce dernier suffise pour donner un bon score à une observation ne correspondant pas au modèle, d'où confusion entre deux éléments du vocabulaire. De plus une étude sur les répartitions des durées des états montre que celles-ci varient peu entre deux occurences d'un même mot. On ajoute donc à chaque état i du modèle une densité de probabilité p(di) de rester une durée di dans cet état, ce qui permet de pénaliser les chemins dans le graphe des états qui s'éloignent par trop de la moyenne. On introduit pour cela des probabilités sur les variables"forward" et "backward" a(i) et b(i) :

at(i)=P(Ol,...Ot,si se termine à t / l)

bt(i)=P(Ot+1,...OT,si se termine à t / l)

On définit également les variables at*(i) et bt*(i) telles que :

at*(i)=P(Ol,..:Ot,si débute à t / l)

bt*(i)=P(Ot+1,...OT,si débute à t / l)

Ces contraintes sur les durées permettent une amélioration du taux de reconnaissance en mots isolés de 90,5% à 93% pour le premier candidat (le meilleur score est obtenu par le mot à reconnaître) et de 98% à 98,8% pour les trois premiers candidats (le mot à reconnaître est parmi les trois premiers candidats).

Lorsque le système de reconnaissance utilise des modèles de mots comme ici pour la reconnaissance de mots connectés trois algorithmes peuvent être mis en œuvre pour concaténer les modèles : le "one-pass" (Bridle 83; Ney 84), le "level-building" (Myers 81) et le "two-level" (Sakoe 79). Les deux derniers nécessitent plusieurs balayages du signal, le premier est en une-passe et est peu coûteux en temps de calcul mais il est difficile de gérer toutes les alternatives de solutions de manière optimale. C'est cet algorithme que nous avons malgré tout choisi d'implémenter en améliorant la gestion des hypothèses [Yé 90]. L'idée de base est de traiter tous les modèles candidats de manière synchrone : une trame de signal est comparée à toutes les références avant l'arrivée de la trame suivante. Cela permet de progresser en parallèle dans tous les modèles (connectés entre eux sous contraintes linguistiques) et de trouver le chemin optimal (fig.3)

Les résultats en mots connectés tombent à 85% pour la première hypothèse et à 94% pour les trois meilleures.

Les modèles HMM ainsi définis sont ensuite intégrés dans une grammaire sous forme d'ATN (§5), c'est à dire que les différentes possibilités de chaînage des mots entre eux sont décrites par un arbre dont chaque noeud pointe vers la chaîne de Markov correspondante. Les résultats obtenus par ce système montrent alors un taux de reconnaissance de 91,85% pour le premier candidat et de 99% pour les trois premiers candidats. Il est intégré à une application de démonstration de dialogue multimodal (ICPdraw).

5. L'ANALYSE LINGUISTIQUE

On sait [Woods, 78] que les ATN (Augmented Transition Networks) sont les outils parmi les plus puissants pour résoudre les problèmes posés par la représentation et la mise en oeuvre des connaissances syntaxico-sémantiques [Sabah, 88]. En effet ils supportent diverses caractérisations syntaxiques et s'accomodent très bien des contraintes sémantiques. On sait cependant que la mise en oeuvre de ces ATN n'est pas très commode pour un utilisateur non averti; c'est pourquoi, pour faciliter l'écriture de la grammaire et du lexique dans un langage agréable, un compilateur a été développé : à partir d'une représentation externe, il produit un ATN en représentation interne muni de relations avec le lexique, lorsqu'on lui fournit une grammaire quelconque en entrée [Reynier, 89]. Parmi les grammaires disponibles nous avons retenu les grammaires lexicales fonctionnelles (GLF) [Bresnan, 82] qui nous ont semblé les plus adéquates pour le langage naturel restreint.

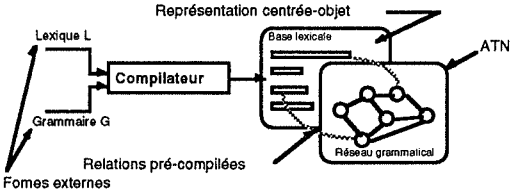


Fig. 4 : Le compilateur du lexique et de la grammaire.

L'analyse syntaxico-sémantique est toujours liée au lexique selon les deux perspectives suivantes:

- la vérification — où il s'agit de confirmer ou d'infirmer qu'une suite de mots est syntaxiquement correcte— qui exige un accès au lexique pour rechercher les attributs syntaxico-sémantiques des mots à vérifier,

- la prédiction (très importante ici) — où il s'agit de

fournir une liste de candidats-mots possibles après une séquence correcte — qui exige aussi un accès au lexique à partir des attributs syntaxico-sémantiques prédits par l'analyseur syntaxique (en examinant cette fois tous les chemins possibles dans l'ATN à partir d'un état origine donné).

Dans tous les cas, il est évident que la relation syntaxe-sémantique-lexique est très forte et doit être prévue au moment de la compilation de l'ATN afin de diminuer, entre autres, le temps de la recherche (une des techniques est de prévoir par avance un accès "statique" précompilé). Cette relation est prise en compte, sous forme de contrainte d'accès, directement dans le lexique — elle porte sur les catégories syntaxiques, sémantiques, attributs, etc. Il est évident que des "actions" placées en partie droite de règles — comme il est de coutume dans les ATN — pourraient résoudre le même problème, mais le temps d'exécution serait plus long puisque les accès seraient calculés à chaque fois : nous avons préféré les traiter par des pointeurs donnant un accès direct aux items lexicaux. Ces contraintes d'accès sont indiquées, si nécessaire, explicitement, dans les règles immédiatement après chaque terme.

Exemple de règle en langage externe :

```
S1:   SN -> Dét N /,accord($1,$2),/;
      /* action d'accord entre le $1=Dét et le $2=N */
S2:   SN -> Dét N Adj($qual);
      /* accès restreint aux adjectifs qualificatifs */
```

Dans la représentation externe, les règles classiques de réécriture sont assorties de deux champs supplémentaires: un champ "contexte" et un champ "actions".

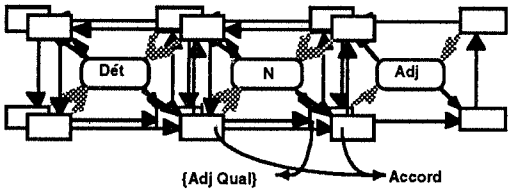


Fig. 5 : Réseau ATN compilé pour les deux règles S1 et S2. Les chaînages entre noeuds sont prévus pour permettre une analyse de droite à gauche et inversement. Un pointeur spécifique est créé pour accéder aux adjectifs qualificatifs (Adj Qual) dans le lexique et une action (Accord) est instanciée à la fois au niveau (Dét N) et au niveau (N Adj).

Une analyse va consister à parcourir le réseau selon le type de fonctionnement fixé par le contrôleur de réseau appelé "analyseur". Deux modes d'analyse sont prévus: le mode ascendant et le mode descendant. Pour chaque mode et à tout moment de l'analyse, deux fonctionnements sont possibles: (a) un fonctionnement en vérification et (b) un fonctionnement en prédiction. Pour vérifier une chaîne d'entrée, l'analyseur cherche un chemin dans le réseau à partir du noeud courant. En prédiction, l'analyseur propose tous les noeuds possibles, successeurs à la distance k, du noeud courant. A partir de points d'ancrage syntaxiques, comme les débuts ou fins de phrase, de syntagme, l'analyse descendante est bien appropriée. Par contre, si l'analyseur ne connait pas la position syntaxique courante mais s'il connait un point d'ancrage du niveau de description du vocabulaire terminal, l'analyse ascendante sera activée. L'analyseur autorise les sens de parcours gauche-droite et droite-gauche. Ainsi, l'analyseur pourra activer une analyse du milieu vers les côtés en partant des points d'ancrage.

L'analyseur permet la gestion des règles récursives. Il maintient deux piles, l'une pour les règles à contexte libre et l'autre pour les règles contextuelles. Il construit en parallèle toutes les solutions syntaxiquement et sémantiquement correctes. En fin d'analyse, il fournit un arbre de solutions syntaxiquement et sémantiquement correctes, la structure des constituants (c-structure), ainsi qu'une liste de solutions fonctionnelles (f-structure).

6. LE FILTRAGE PAR VERIFICATION

La partie vérification que nous proposons est divisée en deux modules fondamentaux :
1- l'apprentissage supervisé par un expert pour constituer un dictionnaire phonétique significatif,
2- la vérification par programmation dynamique (DTW).

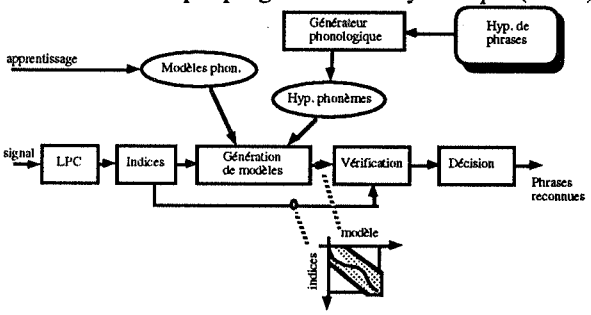


Fig. 6 : Schéma modulaire de la méthode descendante.

L'analyse acoustique contient les mêmes traitements que dans le DAL (§4), on y adjoint un module de calcul d'indices [Caelen, 81] réputé robuste pour la discrimination des phonèmes.
1- La phase d'apprentissage :
Elle consiste à obtenir des informations phonétiques robustes sur tous les mots du vocabualire utile. Pour cela on part du même corpus que celui du DAL, on y marque les phonèmes ou les phases acoustiques réputés stables (peu variants) et peu sensibles au contexte (frictions, occlusions, voyelles longues, etc.) et on fait des statistiques sur ces éléments. On obtient alors un dictionnaire de références phonétiques. On part de l'hypothèse que la vérification reposera sur ces seuls éléments considérés comme pertinents. Pour éviter la

fastidieuse opération de segmentation et d'étiquetage, toutes les répétitions identiques du même mot dans le corpus sont traitées par alignement temporel sur la première grâce à une procédure de programmation dynamique [Rabiner 78], [Sakoe 79] complétée par un suivi de chemin [Chamberlain 83], [Blomberg 88], — technique appelée aussi Time Alignement. Tous les éléments pertinents sont ensuite recueillis dans une base de données phonétiques (modèles phonétiques) qui contient pour chacun sa dispersion multidimensionnelle (en terme d'indices, d'énergie, de durée). Il peut y avoir plusieurs éléments significatifs pour un phonème (typiquement 2 pour les voyelles nasales, 2 pour les occlusives sourdes ou même aucun pour les liquides).

2- La vérification :

Le module de vérification procède par analyse-synthèse : on génère tout d'abord des séquences acoustiques (indices, énergies, durées) compatibles avec les hypothèses de phrase — (re)traduites en chaînes de phonèmes — auxquelles on a associé des valeurs standards prises dans la base de données phonétiques, puis on les compare à la suite des valeurs extraites du signal à vérifier.

Le lexique contient les transcriptions phonétiques (toutes les variantes) de chaque mot, ainsi que les liaisons latentes. Il est donc facile de générer une suite de phonèmes en partant d'une list de mots. Des règles phonologiques sont nécessaires cependant pour obtenir cette séquence de phonèmes correspondant à l'émission d'une suite de mots enchaînés. Dans une première version, le transducteur phonologique filtre le dictionnaire de transcriptions phonétiques de chaque mot de vocabulaire et inclut éventuellement les liaisons entre mots. Nous envisageons dans avenir proche d'installer un transducteur plus complexe.

La vérification proprement dite se fait par comparaison dynamique entre les valeurs du signal à vérifier et les valeurs des références générées, en utilisant des coefficients pondérés selon la pertinence de chaque entité phonétique (par exemple, pour distinguer "déplace" et "dessine" la différence entre /p/ et /s/ est déterminante), et selon chaque indice (dans l'exemple précédent les indices aigu/grave et doux/strident seront plus importants que les autres); la distance entre deux entités phonétiques (f_r, f) est alors :

$$d(f_r,f) = P_r \sum_{i=\text{indice}} p_i\, d_i(I_r,I)$$

où P_r est le coefficient de pondération pour l'entité f_r considérée, p_i est le coefficient de pondération de l'indice i, I_r et I les valeurs d'indices pour la référence et le signal. d_i est une distance statistique (Levenstein) qui tient compte des distributions des indices.

La distance entre deux formes F={f} et F_r={f_r} est finalement $d(F_r,F) = DTW(F_r,F)$.

La décision qui est théoriquement la partie la plus délicate du système de vérification se résume ici à rejeter une hypothèse trop distante de la référence. Dans la mesure où l'on souhaite fournir plusieurs solutions à un module de dialogue (ces solutions sont assorties d'une valeur de distance) aucune pénalisation n'est opérée a posteriori.

Dans l'état actuel de l'expertise, l'apport de ce module de vérification apporte des améliorations sensibles dans le rejet (on sait qu'il est très difficile d'en proposer un avec un modèle markovien seul) et permet de conforter certaines hypothèses ("blanc" contre "bleu", "dessine" contre "déplace"). Il permet également d'opposer "le" et "un" dans des phrases comme "détruis le cercle" et "détruis un cercle" dans lesquelles l'article dénote une opposition importante. Cependant le module phonologique est encore un peu frustre à l'heure actuelle, il ne permet pas de régler le problème de la coarticulation, ce qui oblige à multiplier les références phonétiques candidates à la vérification.

7. CONCLUSION

Les bases d'un système de reconnaissance hybride, modulaire et temps réel ont été définies. Ce système évalué sur une station de travail classique est en cours de portage sur la carte vocale dans l'environnement CPOS développé à l'IRCAM. Il nécessite une puissance de 35 MIPS.

Il comprend trois modules séparables : l'analyseur markovien en mots enchaînés, le contrôleur linguistique et le filtrage par analyse-synthèse. Cela permet d'en intégrer tout ou partie dans un système de dialogue homme-machine à composante vocale selojn la puissance de calcul disponible en ligne.

Par rapport à ses concurrents il offre l'avantage de pouvoir fonctionner en vérification et d'accepter des contraintes linguistiques dynamiques (avec un formalisme de grammaires lexicales fonctionnelles). Par ailleurs, la puissance de la carte vocale sa structure et son environnement de développement permettent d'envisager des algorithmes de reconnaissance (et de compréhension) encore plus évolués.

Remerciements

Ce travail a reçu le soutien financier de la CEE (projet ESPRIT II n° 2105). Nous tenons à remercier tous nos partenaires et plus spécialement L. Sauter, coordinateur du WP "voice".

8. REFERENCES

[Bailly 90] BAILLY G. "Speech Synthesis", Rapport

final, Multiworks n°2105, 1990.

[Baum 67] BAUM L. E. & EGON J. A. "An Inequality with Applications to Statistical Estimation for Probabilistic Functions of a Markov Proces and to Model for Echology." Bull. Amer. Meteorol. Soc.,Vol. 73.

[Blomberg 88] BLOMBERG M."Word Recognition using Synthesized Templates", Speech Transmission Laboratory, Royal Institute of Technology, Stockholm QPSR N2,3 pp.69-81, 1988

[Bresnan 82] BRESNAN J, KAPLAN R.M. "Introduction: Grammar as mental representations of language". *The mental representations of grammatical relations.* in Bresnan ed., Cambridge, 1982.

[Bridle 83] BRIDLE J.S., BROWN M.D. & CHAMBERLAIN R.M. Continuous Connected Word Recognition Using Whole Word Templates. The Radio and Electronic Engineer 53, 167-175, 1983

[Caelen, 81] CAELN J., CAELEN-HAUMONT G. Indices et rporiétés dans la projet ARIAL II. Séminaire DAP, PRC "CHM", Toulouse, 1983.

[Chamberlain 83] CHAMBERLAIN R.M. BRIDLE J.S. "ZIP : a Dynamic Program Algorithm for Time Alignement; two indefinitly long Utterance " Proc. IEEE-ICASSP 83, Boston pp 816-819, 1983.

[Crystal 82] CRYSTAL T.H. & HOUSE A.S. Segmental duration in connected speech signals : Preliminary results. J. Acoust. Soc. Am. 72, 705-716, 1982.

[Delemar 91] DELEMAR O. "Reconnaissance de mots enchaînes par une méthode hybride : reseaux markovien et base de règles" Rapport de DEA SIP, Grenoble., 1991.

[Gagnoulet 89] GAGNOULET C. & JOUVET D. "Reconnaissance de la Parole et Modélisation Statistique : Experience du CNET", l'Echo des Recherches, No 135, 1989.

[Godin 89] GODIN C. & LOCKWOOD P. DTW Schemes for Continuous Speech Recognition : a unified view. Computer Speech and Language 3, 169-198.,1989.

[Haton 91] HATON J.P. PIERREL J.M. PERENNOU G. CAELEN J. GAUVAIN J.L. "Reconnaissance automatique de la parole" DUNOD informatique, Paris, 1991.

[Höhne 83] HOHNE H. COKER C. LEVINSON E. RABINER L.R. "On Temporal Alignement of Natural and Synthetic Speech " IEEE Trans. on Acoust., Speech and Sig. Proc., Vol. 51-6, pp 807-813, Aug. 83.

[Hunt] HUNT M. "Time alignement of Natural speech to Synthetic Speech" Proc.IEEE-ICASSP 84, p 2.5.1.

[Levinson 86] LEVINSON S. E. "Continuously Variable Duration HMM for Speech Analysis", Proc. ICASSP 86, p 1241, 1986.

[Myers 81] MYERS C. & RABINER L.R. A Level Building Dynamic Time Warping Algorithm for Connected Word Recognition. IEEE Trans. ASSP-29, 284-297, 1981

[Ney 84] NEY H. "The Use of an One Stage Dynamic Programming Algorithm for Connected Word Recognition", IEEE Trans, ASSP-33, pp 263-271, 1984.

[Poritz 88] PORITZ A. B. "Hidden MARKOV Models : a Guided Tour", IEEE Trans 88, pp 7-13, 1988.

[Rabiner 89] RABINER L. R. "A Tutorial on HMM and Selected Application in Speech Recognition", Proc. of the IEEE, Vol 77, No 2, pp 257-286, Feb. 1989.

[Reynier 90] REYNIER E. Syntaxe et sémantique dans le système DIRA. Thèse de l'INPG, Grenoble, 1990.
[Sakoe 79]

SAKOE H. "Two Level DP-Matching. A Dynamic programming Based Pattern Mathcing Algorithm for Connected Word Recognition", IEEE Trans. on Acoust., Speech and Sig. Proc., Vol. 27-6, pp 588-595, Dec. 1979.

[Viterbi 67] VITERBI A. J. "Error Bounds for Convolutional Codes and an Asymptotically Optimal Decoding Algorithm.", IEEE Trans.Informat. Theory, Vol. II-13.

[Wagner 81] WAGNER M. "Automatic Labelling of Continous Speech with a given Phonétic Transcription using Dynamic Programming Algorithms ", proc.IEEE-ICASSP 1981, p 1156.

[Ye 90] YE H. & CAELEN J. Duration Constraints for Speech Input Interface in the MULTIWORKS Project. International Conference on Spoken Language Processing. Nov. 18-22 1990, Kobe, Japan.

19es J.E.P. - Bruxelles, 19 au 22 mai 1992.

UN SYSTEME DE DIALOGUE MULTIMODAL POUR POSTE DE TRAVAIL INTELLIGENT FONDE SUR UNE GRAMMAIRE LEXICALE FONCTIONNELLE.

Fabrice DUERMAEL ; Jean-Marie PIERREL

CRIN-CNRS & INRIA-Lorraine
BP 239 - F54506 Vandoeuvre-lès-Nancy cedex

Résumé

Nous étudions dans cet article l'interface avec un opérateur travaillant sur un poste de travail intelligent, en nous focalisant plus particulièrement sur l'utilisation simultanée de la parole et de médias de désignation. Un tel système est destiné à assister un opérateur sur console dans des tâches de supervision ou de commandement. En partant de l'exemple du démonstrateur multimodal MELODIA, nous montrons que cela nécessite une bonne prise en compte de la situation de communication à tous les niveaux de processus de compréhension de messages. Alors que les systèmes de dialogue typiques attendent des composants ayant le plus de visibilité à corriger les interprétations erronées effectuées par ceux qui en ont le moins, on est ici amené à introduire de nouvelles interactions dès les niveaux syntaxiques afin de mieux prendre en compte le contexte global. Cela nous amène finalement à discuter de l'opportunité d'une composante pragmatique dans un analyseur fondé sur une grammaire lexicale fonctionnelle.

Le système présenté fait l'objet d'une convention d'étude entre THOMSON-CSF[1] et l'équipe DIALOGUE du CRIN-CNRS et de l'INRIA-Lorraine(Pouteau, 1990 ; Duermael, 1991). Elle concerne la mise en œuvre d'un démonstrateur multimodal (MELODIA[2]) sur station de travail La particularité d'un tel système est de reposer sur une station de travail utilisant différentes modalités de communication (la parole, le geste, l'affichage graphique) comme poste de travail dans un contexte opératif. Il est important de noter qu'en tant que démonstrateur, ce système n'est pas destiné à devenir un produit fini mais au contraire une plateforme pour effectuer des tests partiels de validité. Les applications concernées sont de type supervision de processus ou élaboration de directives. Dans le cas présent, l'application retenue est le contrôle de trafic aérien (ATC).

L'objectif de cet article est double. En premier lieu, il s'agit de montrer quels sont les prérequis spécifiques d'une interface pour poste de travail intelligent (PTI) par rapport à 1) les systèmes de dialogue oral homme/machine classiques ; 2) les interfaces graphiques homme/machine. Dans un PTI, le but se conçoit bien moins en privilégiant telle ou telle modalité de communication qu'en les faisant cohabiter et en tirant parti de leurs spécificités : c'est ce que l'on entend de manière générale par système multimodal.

Tout en restant lié à ces prérequis, nous nous attacherons ensuite plus particulièrement aux problèmes relatifs à la compréhension d'énoncés multimodaux où intervient une composante orale. Dans les systèmes de dialogue typiques, l'énoncé est traduit en pré-commandes dès l'analyse syntaxico-sémantique sans quelque accès à un contexte linguistique ou extralinguistique. Cela a pour effet de reléguer toutes les opérations utilisant le contexte à la gestion du dialogue. Le système présenté (MELODIA) occupe une position intermédiaire. Sa partie "compréhension d'énoncés" repose sur une grammaire lexicale fonctionnelle qu'il a fallu étendre pour prendre en compte une partie de la situation de communication, ce qui s'avère primordial pour un analyseur multimodal.

Néanmoins, les choix effectués dans MELODIA ont été faits en toute connaissance des limites que cela imposait

[1]Service PTI (Poste de Travail Intelligent) de la direction technique de la division SDC (Systèmes de Détection et de Contrôle).

[2]MELODIA : Multimodal Environment for a naturaL and task-Oriented DIAlogue.

au système. Si l'on désire aller plus loin, il faudra non seulement mieux prendre en compte le contexte global mais effectuer des processus d'inférences sans quoi on ne pourrait à proprement parler de compréhension. La problématique se pose alors en ces termes : vaut-il mieux étendre encore les mécanismes d'unification déjà sophistiqués dans une grammaire lexicale ou au contraire revenir à une analyse structurelle de syntagmes en la faisant collaborer avec une sémantique mieux adaptée.

LES IMPERATIFS D'UN POSTE DE TRAVAIL INTELLIGENT

L'intervention d'un opérateur dans des tâches automatisées relève de la supervision de systèmes (le contrôle) et/ou de l'éboration de directives (la commande). C'est néanmoins lorsque la marche normale du système se dégrade que l'opérateur est successivement amené à évaluer l'état d'un système dont il ne connaît pas tous les paramètres, à déterminer une stratégie de résolution et, à partir de là, prendre les décisions adéquates, parfois lourdes de conséquences. Cette tâche est rendue d'autant plus ardue que le degré d'automatisation atteint est élevé. Dans un tel contexte, l'opérateur tend à jouer un rôle plus ou moins passif en tant que superviseur du système mais il peut devenir brusquement sollicité lorsqu'il rentre dans des phases de décision ou en cas de situation critique.

Pour pouvoir prétendre à une aide efficace de la part du système informatisé, il est nécessaire de prendre en compte des facteurs concernant l'utilisateur pour mieux le suppléer ou le seconder dans un contexte opérationnel fluctuant. Dans de telles conditions, il s'avère primordial de modéliser tant la tâche et les connaissances de opérateur que ses modes de communication. Cela constitue autant de représentations de connaissances qui rentreront en jeu dans l'interface entre l'opérateur et l'application automatisée.

POURQUOI UNE INTERFACE "INTELLIGENTE" ?

Les signaux reçus par les médias de perception de l'interface reflètent toute l'activité mentale de l'opérateur, qui a su planifier une série d'actions, se les représenter sous forme de concepts pour finalement les exprimer. Pour que l'interface puisse prétendre retrouver les intentions de l'opérateur, elle doit être en mesure de parcourir le chemin inverse : c'est ainsi que l'on peut définir la faculté de compréhension dont elle doit faire preuve. Heureusement, le caractère finalisé de l'interaction permet de simplifier grandement ce processus de compréhension.

LA FINALITE DE L'INTERACTION HOMME/MACHINE

Nous nous plaçons dans le cadre d'un dialogue finalisé, où le but de l'utilisateur est de communiquer avec une application informatique en vue de réaliser un certain nombre de tâches, planifiées de manière implicite ou explicite. D'une manière générale, les sous-buts sont multiples et reliés entre eux. C'est en se restreignant de la sorte que l'on peut espérer se limiter dans un domaine bien défini. La finalité apparaît même comme la condition *sine qua non* de représentation des connaissances, tant au niveau linguistique que de l'application.

Nous allons ainsi nous limiter au domaine du contrôle de trafic aérien. Le dialogue entre le contrôleur et la machine se résume de la façon suivante (Bacconnet, 1991) : accepter et assumer un vol, déplacer un strip[3], changer son format, signaler des conflits, retirer un vol de manière anticipée.

LA COMPREHENSION D'ENONCES MULTIMODAUX : PLACE DES SYSTEMES SEMIOTIQUES

Concernant la communication linguistique, F. Rastier (1989, p. 50) écrit :

"Une ou plusieurs sémiotiques associées [à la parole] sont toujours présentes ; et l'on pourrait dire que la communication linguistique est de nature plurisémiotique.":

Cela est aussi retenu en pratique dans les systèmes multimodaux (Cf MMI2, MELODIA) : la parole y est considérée comme le pivot du dialogue. Insistons néanmoins sur le fait que cela ne concerne que la communication linguistique.

Or, des systèmes sémiotiques[4] (Cf E. Benveniste, 1969, pp. 51-54) de type différent ont largement fait leur preuve dans le domaine des interfaces homme-machine. Il s'agit notamment des signes qui à un symbole graphique (une icône) associent une action. C'est sur un tel système[5] que repose le paradigme de la manipulation directe d'objets graphiques. On retrouve des fonctions identiques associées aux signes dans les menus déroulants (une action est associée à un mot).

E. Benveniste définit la notion de convertibilité entre systèmes sémiotiques en partant du principe de leur non redondance : deux systèmes sont convertibles s'ils

---

3Dans le jargon des contrôleurs aériens, un strip est une bande de papier (ou sa représentation sur écran) sur laquelle figurent les caractéristiques d'un plan de vol (niveaux de vol, horaires de passage devant des balises, etc...).

4Un système sémiotique est caractérisé par :
- son mode opératoire (i.e. la manière dont le système agit, notamment le sens auquel il s'adresse)
- son domaine de validité (i.e. les circonstances où il doit être reconnu)
- la nature et le nombre de ses signes
- son type de fonctionnement (i.e. la relation qui unit les signes et leur confère fonction distinctive)

5Citons un système sémiotique du même type rencontré dans la vie quotidienne : celui des feux du trafic routier.

possèdent même rapport de signification (*ibid.*, p. 53). Par exemple, les systèmes où une action est signifiée par une icône ne sont pas convertibles en un système linguistique. Autrement dit, la valeur d'un signe ne se définit seulement qu'à l'intérieur de son système, par l'intermédiaire des différences fonctionnelles avec les autres signes.

Pour résumer, nous pouvons dire que nous avons en présence deux grandes classes de systèmes sémiotiques dans les interfaces multimodales :
• les systèmes dont le signifié d'un signe (un graphème, une icône) est directement et seulement une action. Le rôle d'un média de pointage est alors d'activer l'inférence inhérente au signe dans une situation de communication (Cf Eco, 1988). C'est également le cas pour des combinaisons de touches au clavier.
• le système relatif à la communication linguistique, lui même plurisémiotique. Dans ce système, la langue est le pivot en apportant une cohésion au message[6]. Nous allons nous focaliser plus particulièrement sur ce mode de communication dans la suite de ce chapitre.

N.B. Un média peut rentrer indifféremment dans l'une ou l'autre de ces classes. Chacune possède ses avantages propres : l'expression de directives complexes ou d'intentions pour l'une et la commodité d'utilisation pour l'autre (Cf J. Coutaz, 1990, chapitre 2 : les modèles GOMS et Keystroke). Les deux rentrent toutefois en jeu dans l'interaction avec le déroulement de l'application.

Ce sont tous ces prérequis qui ont guidé nos choix dans MELODIA. Nous allons préciser et discuter ses caractéristiques compte tenu des limites que nous nous étions imposé.

CARACTERISTIQUES DE MELODIA : LES LIMITES IMPOSEES

MELODIA repose sur un module de gestion du dialogue et de la tâche situé au-dessus d'un module de compréhension de messages multimodaux. Il s'agit d'un choix qui a été effectué pour faciliter une mise en œuvre incrémentale. Son principal défaut est de poser de manière arbitraire une hiérarchie injustifiée entre des modules.

GESTION DU DIALOGUE ET DE LA TACHE

Le dialogue étant orienté par la tâche, nous avons choisi d'effectuer conjointement dans un même module la gestion du dialogue et la planification de la tâche. Nous avons opté dans un premier temps pour un dialogue rigide.

Les tâches sont représentées selon le formalisme MAD[7] (Pierret, Delouis, Scapin, 1989). Il consiste à représenter pour chaque tâche-objet le but de la tâche, les paramètres d'entrée et les préconditions qu'ils doivent vérifier, le corps de la tâche (composition d'atomes et de sous-tâches), les résultats de sortie et les postconditions qu'ils doivent vérifier. En liaison avec la gestion de la tâche, le module de dialogue se comporte comme un automate à états finis, dont les états sont les suivants (Cf Bacconnet, 1991) :
• attente d'événements : aucun dialogue n'est en cours ;
• attente de paramètres : attente de réponse lorsque certains paramètres sont manquants ;
• attente de prédicat ;
• attente de conditions : certaines conditions sur des paramètres n'ont pu être vérifiées.

LA COMPREHENSION DE MESSAGES MULTIMODAUX

Le module correspondant n'est destiné à traiter que les messages multimodaux dont le langage oral est prédominant. La manipulation directe d'objets, on l'a vu, peut permettre d'exécuter des actions équivalentes à un message vocal auquel sont associées des désignations[8]. Pour assurer une cohérence dans le dialogue, il est nécessaire que l'un comme l'autre, par quelque moyen que ce soit, aient les mêmes effets sur la représentation dans l'interface des objets de la tâche.

Reconnaissance de la parole et problèmes associés
Le traitement de la parole dans MELODIA repose sur une reconnaissance globale[9] qui nécessite la donnée d'une grammaire formelle à contexte libre. Même si cela nous assure d'une construction syntaxique relativement correcte, cela présente le gros désavantage de fournir au module de compréhension l'expression de l'opérateur phrase après phrase en toute ignorance de contexte. Cela oblige donc à répertorier tous les énoncés elliptiques qu'il est capable de produire dans toutes les situations de communication envisageables. Dans ces conditions, le problème est qu'en tolérant à peu près n'importe quoi indépendamment du contexte, la fiabilité de la reconnaissance en souffre beaucoup. Nous nous limiterons pour cette raison à des reprises elliptiques élémentaires. Le système PARTNER (Cf Morin, Pierrel, 1987 § 3.2.2.b) avait partiellement résolu ce problème particulier en sélectionnant selon l'état du dialogue une sous-grammaire sémantique. En cas de mauvaise reconnaissance (par exemple lorsque l'opérateur ne répond pas à la question posée), il était possible de relancer la reconnaissance sur le signal bufferisé avec la grammaire générale.

---

[6] En plus de sa valeur locutoire, le message peut avoir une valeur illocutoire ou perlocutoire (selon la terminologie austinienne).

[7] Méthode Analytique de Décomposition des tâches.

[8] Par l'intermédiaire d'une souris ou d'un gant DATAGLOVE de VPL.

[9] Système DATAVOX de VECSYS.

Le module de reconnaissance vocale fournit ainsi au module de compréhension trois phrases potentielles. La traduction en une forme structurelle logique destinée au module de dialogue s'effectue par le biais d'une analyse syntaxico-sémantique reposant sur la théorie lexicale-fonctionnelle de Bresnan et Kaplan (1981). Elle appartient à l'ensemble des *grammaires à base de traits attribut-valeur*, qui utilisent toutes une représentation syntaxique des énoncés sous la forme d'ensembles imbriqués de structure attribut/valeur. Le mécanisme privilégié de résolution étant celui de l'*unification de traits*, on confond souvent cet ensemble avec celui des *grammaires à base d'unification*. Chaque structure est en fait l'annotation (ou étiquette) d'un syntagme de la phrase.

La théorie lexicale-fonctionnelle
Il est important de noter dans ces formalismes que, même si les structures attribut/valeur y sont prépondérantes, une représentation arborescente des constituants syntagmatiques est presque toujours sous-jacente. Néanmoins, sa part est toujours réduite en imposant aux syntagmes des contraintes relativement faibles quant aux places qu'ils doivent occuper. Pour une étude approfondie de tels formalismes, on pourra se reporter à l'ouvrage de Johnson (1988), où se trouvent décrites les propriétés logiques (consistance, complétude, décidabilité) qui leur sont communes. Du fait de leur facilité d'implémentation, elles offrent dans le champ de la syntaxe formelle une alternative intéressante aux théories structuralistes plus classiques.

Nous avons vu que les structures attribut-valeur (dites dans les GLF structures fonctionnelles) devaient être considérées en tant qu'étiquette d'un syntagme. Il est utile de rappeler ici ce qu'on entend par syntagme. D'après Grevisse (4°), un syntagme est un groupe de mots formant une unité dans une proposition (ou sous-phrase, phrase, ...). Cette unité est définie par des critères distributionnels (Cf Benveniste E., 1962, p 120). Ils sont caractérisés par leur fonction dans la proposition et contiennent chacun un noyau et son périphérique.

Un attribut dans une structure fonctionnelle peut ainsi revêtir différents usages :
• Indiquer un trait syntaxique associé à un syntagme ;
• Indiquer le noyau (PREDICATE) ou la partie périphérique (XADJUNCT) associé à un syntagme ;
• Préciser la fonction (rôle) du syntagme associé (le gouverneur) par rapport à des syntagmes se trouvant à une hiérarchie inférieure (les gouvernés)[10].

[10]On notera à propos de ce point la différence de point de vue concernant la fonction associée à un syntagme. En fait, on doit se placer selon le gouverné et non le gouverneur ; ce qui revient à dire que l'on recherche la portion d'énoncé qui englobe le syntagme en question.

Aspects sémantiques
Les structures fonctionnelles sont destinées à servir d'entrée au composant sémantique. Pour assurer le lien, les entrées lexicales spécifient une correspondance directe entre arguments sémantiques et les configurations des fonctions grammaticales de surface. Nous avons néanmoins enrichi le mécanisme d'unification pour ne la permettre que si certaines propriétés directement liées au référent sont vérifiées. Il s'agit donc d'une extension véricconditionnelle de l'analyseur. Cela suppose une représentation des propriétés des référents dans le lexique de l'analyseur syntaxico-sémantique, ce qui revient à poser les caractéristiques du seul monde possible de référence. Aussi discutable que cela puisse paraître, nous pouvons toutefois l'admettre dans le cadre de langages artificiels de commande fortement restreints où nous pouvons nous contenter d'associer un état du monde à un énoncé. Cela passe par la détermination des référents, qui peut se faire soit par désignation directe, soit par reprise anaphorique.

Désignation directe et reprise anaphorique
Bien que cela soit réducteur, nous avons choisi de n'effectuer des désignations directes qu'avec des pronoms démonstratifs et des reprises anaphoriques qu'avec des pronoms ou groupes nominaux définis. En effet, il est tout aussi possible d'effectuer une désignation directe avec un groupe nominal défini ou une anaphore avec un adjectif démonstratif. L'avantage de cette restriction est alors d'apporter une solution simple pour traiter de la communication multimodale synchrone. En rajoutant au niveau des syntagmes de certaines règles de réécriture grammaticales la nécessité de l'accompagnement d'une désignation. C'est donc seulement avec une telle restriction que la prise en compte de la désignation peut intervenir à un niveau syntaxique de manière pertinente. Notons toutefois qu'il s'agit là d'un début de prise en compte de la situation de communication à un niveau syntaxique.

Mais si l'on désire aller au-delà, il est nécessaire de faire cohabiter des médias où l'accès à la référence est différent : il est direct dans le cas de la désignation alors que dans le cas de la parole, il est déterminé par des relations sémantiques dans la situation courante de communication.

Tout cela nous conduit à nous interroger sur la place de ce qui est traditionnellement réservé à la pragmatique - à savoir la prise en compte de la situation de communication.

LA PLACE DE LA PRAGMATIQUE EN QUESTION

Classiquement et de manière fort discutable, les niveaux syntaxe et sémantique font partie du module de compréhension et le niveau pragmatique de celui du dialogue. Dans ces conditions, le rôle de la pragmatique est de lier les anaphores aux référents, analyser la structure du discours et vérifier la consistance du sens deviné avec les connaissances sur la tâche dans la

situation courante de communication (Cf Ait-Kaci, 1989).

Nous partageons la conviction que certains aspects traditionnellement relégués au niveau dialogue et gestion de la tâche sont fondamentaux pour une meilleure prise en compte de l'énoncé au niveau compréhension. Cela est encore plus important en communication multimodale, où les modalités ne doivent pas être traitées indépendamment.

Dans MELODIA, seulement une partie de ces points ont été retenus dans le module de compréhension. La prise en compte du contexte (sous forme d'historique du dialogue) permet de résoudre certains cas d'ellipses. La détermination des référents utilise une représentation de l'état de l'interface et un historique linéaire du dialogue. Une partie des connaissances sur la tâche sont stockés sous forme de traits référentiels. Enfin, l'analyse syntaxico-sémantique de la parole a accès à une partie du contexte non linguistique, en l'occurrence les objets désignés.

Cela ne veut pas dire que tous les aspects du dialogue doivent être intégrés à un module de compréhension. Ce qui relève de la gestion de la situation de communication revient de plein droit à la pragmatique du dialogue. Nous pouvons prendre et discuter l'exemple d'une grammaire d'unification englobant des contraintes syntaxiques, sémantiques et pragmatiques dans un petit sous-langage de l'anglais (Ait-Kaci, 1989). Remarquer des "accrochages sémantiques" en même temps que l'analyse syntaxique permet certes d'optimiser le traitement mais ne résoud pas le problème de fond pour une utilisation à plus grande échelle.

Un tel système repose comme MELODIA sur une sémantique vériconditionnelle, avec les limites qu'on lui connaît : mélange des connaissances linguistiques et extralinguistiques, pas de critères pour l'exhaustivité des traits dans le lexique. Or plutôt que d'exprimer des connaissances sous forme de traits référentiels, ne vaudrait-il pas mieux avoir utilisé directement des connaissances sur la tâche par l'intermédiaire d'une sémantique mieux adaptée ? Il nous apparaît en effet que même si l'unification est un outil puissant, il ne permet de prendre en compte à lui seul le langage. Dans ces conditions, il est opportun de spécialiser l'unification à la syntaxe de surface et que son objet ne soit plus la phrase, abstraction mal définie, mais l'énoncé en contexte. C'est notamment par cet intermédiaire qu'on pourra espérer prendre en compte correctement les phénomènes d'ellipses.

VERS UNE MEILLEURE PRISE EN COMPTE DE LA SITUATION DE COMMUNICATION

Il est maintenant admis dans le domaine des interfaces homme-machine que l'interface doit disposer de connaissances sur l'application elle-même. On pourrait qualifier ces connaissances comme étant spécialisées dans le domaine d'application concerné. Il apparaît ainsi primordial de considérer les tâches de l'opérateur comme des objets que l'interface peut manipuler et utiliser comme base à des raisonnements. De cette façon, les tâches ne sont plus seulement des procédures informatiques à faire exécuter par l'application.

Pour connaître à un moment donné l'état dans lequel se trouve l'application, il est difficilement envisageable de l'interroger régulièrement. La précision obtenue serait au détriment d'un mécanisme de filtrage important et d'une extrême dépendance vis-à-vis de l'application (Romary, Gaiffe, Pierrel, 1991). C'est pour cela qu'il est préférable de tenir à jour les propriétés des objets par la modélisation des actions, même si l'on perd au niveau précision[11]. C'est également l'approche retenue[12] dans des systèmes de compréhension de textes (Cavazza, 1991), où collaborent complémentairement trois sémantiques : une inférentielle, une différentielle et une référentielle. Cette dernière est une sémantique procédurale, où des procédures associées à des concepts (liés à des syntagmes) donnent une "vie" à un modèle situationnel. On entend par là une représentation plausible de la situation construite à partir de la succession des énoncés et des différentes connaissances - linguistiques, de sens commun, spécialisées - dans le domaine concerné. Ces procédures peuvent correspondre à (cf. *ibid.*, p.92) :

- l'introduction d'objets dans le modèle en précisant leurs propriétés (concepts "arguments" correspondant à certains groupes nominaux) ;
- la simulation des effets des actions sur les objets (concepts "prédicat" correspondant à des actions ou des processus sous forme verbale, nominale ou plus rarement adjectivale).

Vu sous cet angle, la représentation de l'état de l'application (par un modèle situationnel) et la simulation des effets des concepts véhiculés par les énoncés permettent une véritable symbiose entre la gestion du dialogue opérateur/système et la gestion de la tâche. La représentation de la vie des objets à travers leurs différents états, la gestion des propriétés des objets et de la continuité des actions passent par la continuité du discours (Romary, Gaiffe, Pierrel, 1991). C'est notamment cette continuité qui rend de la cohérence aux objets, liant ainsi les énoncés se référant aux mêmes objets. Par le biais d'une composante sémantique inférentielle, un modèle situationnel permet la reconnaissance de scripts. De cette manière, on peut situer l'action courante dans une perspective plus large.

BIBLIOGRAPHIE

---

[11] On distinguera de ce fait les propriétés visibles au niveau du dialogue, et celles qui ne le sont pas.

[12] Sans avoir le choix cette fois-ci, puisqu'il n'y a pas d'interaction avec une application informatique.

AIT-KACI H., LINCOLN P., 1989,
LIFE, a natural language for Natural Language in t.a. informations, volume 30

BACCONNET B., 1991,
Dialogue H/M adaptatif et multimodal pour PTI, Thomson CSF-SDC, Bagneux

BENVENISTE E., 1962,
Les niveaux de l'analyse linguistique in Problèmes de linguistique générale, Tome 1, Coll. TEL, Gallimard, pp 119-131, 1974

BENVENISTE E., 1969,
Sémiologie de la langue in Problèmes de linguistique générale, Tome 2, Coll. TEL, Gallimard, pp 43-66, 1974

BRESNAN J. & KAPLAN R., 1981,
Lexical-functional grammar: a formal system for grammatical representation, in The mental representation of grammatical relations, MIT Press, Cambridge

CAVAZZA M., 1991,
Analyse sémantique du langage naturel par construction de modèles, thèse de doctorat de l'Université de Paris VII

COUTAZ J., 1990,
Interfaces homme-ordinateur, Conception et réalisation - Dunod Informatique

DUERMAEL F., 1991,
Spécification d'une interface opérateur/application, Rapport de DEA, Université de Nancy I

ECO. U., 1984,
Sémiotique et philosophie du langage, Formes sémiotiques, Gallimard, Paris

JOHNSON M., 1988,
Attribute-value logic and the theory of grammar, Center for the study of language and information, Lecture Notes number 16, Stanford

MORIN P., PIERREL J.-M., 1987,
Partner : un système de dialogue homme-machine, Cognitiva 87, Paris

PIERRET C., DELOUIS I., SCAPIN D.L., 1989
Un outil d'acquisition et de représentation des tâches orienté-objet, Rapport de recherche INRIA n° 1063

POUTEAU X., 1990,
Dialogue H/M multimodal pour PTI, Mémoire de DEA, Université de Nancy I, CRIN

RASTIER F.,1989,
Sens et textualité, Coll. Langue, linguistique, communication, Hachette Université, Paris

ROMARY L.,GAIFFE B.,PIERREL J.-M., 1991,
Eléments de convivialité dans un dialogue de commande multimodal, Préactes IHM'91, Dourdan, 11-13 décembre 1991

19es J.E.P. - Bruxelles, 19 au 22 mai 1992.

Prévention et gestion des erreurs de reconnaissance et de compréhension dans un système de dialogue oral.

S.GITTON

SNCF direction de la Recherche - Dept Prospective
45, rue de Londres
75379 PARIS CEDEX 08 - FRANCE

Centre National d'Etudes des Télécommunications
LAA/TSS/RCP - BP 40
22301 LANNION - FRANCE

E-mail : gitton@lannion.cnet.fr

Résumé

Ce papier présente différentes stratégies permettant de prévenir et de gérer les erreurs de reconnaissance et de compréhension dans un système de dialogue oral. La spécification des types de dialogues à modéliser se fonde sur la recherche d'un compromis entre l'efficacité du dialogue, la liberté d'expression dont la prise d'initiative par l'utilisateur est le corollaire et la robustesse aux erreurs. Concernant la prévention des erreurs, la solution consiste à proposer à l'utilisateur un langage lui permettant d'exprimer des initiatives portant aussi bien sur le contrôle de la résolution de son problème, que sur le contrôle d'un processus de gestions des erreurs de dialogue, tout en assurant une reconnaissance suffisante. Le traitement des erreurs par le système, s'appuie quant à lui sur la gestion de contraintes de cohérence du dialogue et des stratégies de confirmations.

INTRODUCTION

La réalisation de serveurs vocaux interactifs (SVI) "grand public", fondée sur la reconnaissance d'une trentaine de mots isolés est en passe de devenir une réalité industrielle, comme semble le montrer par exemple l'expérimentation *'Les Baladins'* [TOULARHOAT 92] menée actuellement par le CNET. Un projet entre le CNET et la SNCF vise la réalisation d'une maquette de SVI "grand public", acceptant de la parole continue en entrée. Le travail présenté ici s'inscrit dans le cadre de ce projet.

La première partie de ce papier met en évidence les principales caractéristiques du système de dialogue mis en oeuvre pour cette maquette, à travers l'étude des contraintes auxquelles il est soumis. La principale de ces contraintes étant d'assurer une forte robustesse aux erreurs, nous présentons dans la seconde partie les solutions envisagées pour prévenir ces erreurs, spécialement dans le contexte d'une prise d'initiative de l'utilisateur. La troisième partie détaille les éléments de détection et de correction d'erreurs permettant d'assurer le meilleur compromis entre l'efficacité du dialogue, la liberté de l'utilisateur et la robustesse du système. Enfin la dernière partie présente l'expérimentation qui permettra de valider les solutions.

1 PRESENTATION DU SYSTEME

L'objectif du projet est de définir, d'étudier et de mettre en place un serveur vocal interactif utilisant la reconnaissance de parole à travers le réseau téléphonique. Ce système doit permettre l'interrogation par le "grand public" d'une base de données contenant les informations sur les horaires de trains (en fonction des gares de départ et d'arrivée, d'une date et d'une heure de départ ou d'arrivée, approximative). Il doit aussi permettre de présenter les tarifs pour chacun de ces horaires en fonction de la classe et d'un taux de réduction. De plus l'accès à un calendrier simple indiquant les périodes d'application des tarifs et une présentation des principales réductions avec leur mode d'application semblent nécessaires. En effet, le choix du taux de réduction dans le calcul du prix est laissé à la responsabilité de l'utilisateur. Il faut, par conséquent, lui donner les moyens d'effectuer ce choix.
Les principales contraintes de réalisation imposées se répartissent en trois catégories : celles liées au mode d'accès à l'information, celles liées aux performances intrinsèques du système de

reconnaissance utilisé et de son implémentation, et, enfin, celles liées à la tâche elle-même.

1.1 Contraintes liées au mode d'accès

Ce serveur doit être accessible au "grand public" au moyen du téléphone. Le système de reconnaissance utilisé doit donc être de type multi-locuteur. Il doit être apte à reconnaître de la parole continue, apparaissant comme naturelle, c'est à dire contraignant au minimum les possibilités d'expression de l'utilisateur. La modélisation du langage de ce dernier, doit donc tenir compte des effets propres à l'utilisation de l'oral comme mode de communication, tels que la présence de reprises ou d'hésitations. Afin de rendre le dialogue entre l'usager et le système efficace et de diminuer le nombre d'erreurs dues à un non respect des consignes, il semble en effet utile d'accepter une certaine liberté dans l'expression des informations données par l'utilisateur. De même, le nombre des instructions (quelques mots ou formules-clés par exemple) qu'il devra apprendre pour manipuler le système doit rester faible et leur emploi simple. Une conséquence de cette liberté apparente, laissée à l'utilisateur risque de se traduire de sa part, par une prise d'initiative fréquente. Autant cette liberté peut être intéressante pour la concision qu'elle peut procurer, autant elle peut totalement dégrader l'efficacité du dialogue si elle n'est pas canalisée. Ceci est dû, en partie, aux limites de reconnaissance actuelles.

1.2 Contraintes matérielles

Le système de reconnaissance utilisé est le logiciel développé au CNET [JOUVET 91] fondé sur la méthode de reconnaissance statistique à partir de modèles de Markov cachés. Il est implanté sur une carte RDP50. La reconnaissance en temps réel nécessite, au préalable, le chargement sur cette carte d'un réseau markovien regroupant les informations syntaxiques, lexicales et acoustiques permettant de modéliser l'ensemble des phrases susceptibles d'être reconnues. La place disponible sur la carte limite le nombre de contraintes syntaxiques pouvant être introduites ainsi que la taille du lexique. L'ensemble "logiciel et carte" était initialement prévu pour une cinquantaine de mots. Un réseau permettant de reconnaître une date de départ ou d'arrivée avec une plage horaire (p.ex. "en début d'après midi") ou une heure (sans les minutes) plus quelques mots-clés nécessaires à la gestion du dialogue, dépasse les cent mots. Le taux de reconnaissance diminue évidemment avec l'augmentation de la taille du vocabulaire mais aussi avec le relâchement des contraintes syntaxiques.

1.3 Contraintes liées à la tâche

Une contrainte irréductible liée à la tâche est que les noms de gares à reconnaître sont au nombre de cinq mille. Il est impossible de les intégrer dans un unique réseau de reconnaissance. Même si leur acquisition fait l'objet d'un traitement particulier (phase "épellation"), ils risquent de perturber fortement la reconnaissance s'ils sont prononcés de façon sur-informative dans les réponses de l'utilisateur, évènement difficile à éviter dès lors que l'on accepte la sur-information par ailleurs.

2 SPECIFICATION DES DIALOGUES

La spécification des dialogues à modéliser se fonde sur la recherche d'un compromis entre l'efficacité du dialogue, la liberté d'expression de l'utilisateur et la robustesse aux erreurs. Ce dernier point apparaît être le point prioritaire. Il nécessite l'emploi de moyens de prévention, de détection et de correction des erreurs.

2.1 Prévention des erreurs

2.1.1 Principes et définitions

La prévention d'erreurs réside dans la mise en oeuvre de techniques visant à réduire la complexité prévisible des interventions de l'interlocuteur de manière à les rendre au moins en partie reconnaissables. Ce problème a aussi été évoqué dans [GUYOMARD 91].

Une intervention est considérée reconnaissable dans deux cas :
(1) si la transcription (supposée unique) sous forme de mots, correspondant à ce qui a été émis, est réalisée en temps réel, à partir du réseau chargé sur la carte au moment de l'émission de l'intervention;
(2) si une transcription partielle correcte est réalisée par le réseau chargé sur la carte et que le reste de la transcription est effectué par un second réseau, à charger sur la carte, déterminé de façon unique à partir de la transcription déjà réalisée et du contexte de l'intervention (l'ensemble de ces deux opérations devant être effectué en un temps acceptable pour l'utilisateur, compte tenu du dialogue).

Une intervention est considérée en partie reconnaissable si une transcription partielle correcte est réalisée par le réseau chargé sur la carte et que le

reste de la transcription est constitué de "jokers" associés uniquement à des portions de l'intervention que l'on sait pertinemment ne pas pouvoir reconnaître (comme p.ex.: les noms de villes exceptée une liste d'une vingtaine de noms).

2.2.2 Les stratégies proposées

La complexité prévisible d'une intervention de l'utilisateur, dépend de quatre facteurs.

Premièrement, elle est fonction du nombre d'actes (informer de la ville de départ, effectuer une confirmation, demander la modification d'un paramètre,...) qu'il est susceptible d'accomplir dans l'étape suivante de sa démarche de résolution du problème. (Pour une formalisation de la notion d'acte de dialogue voir [SADEK 91].) Si l'on se place dans l'hypothèse d'un dialogue complètement dirigé par le système, c'est à dire ne ménageant pas à l'utilisateur la possibilité de prendre l'initiative, ce nombre d'actes est restreint. Dans le cas contraire, des mécanismes doivent être mis en place de manière à diriger l'utilisateur dans l'expression de cette initiative. Ceci est examiné dans la section 2.1.3.

Deuxièmement, la complexité en question dépend du nombre d'actes que l'utilisateur peut fusionner en un tour de parole. Une réduction de cette complexité s'obtient ainsi, d'une part, en diminuant le nombre d'actes émis par le système dans son tour de parole (comme éviter de poser des questions réalisant des confirmations implicites) et, d'autre part, en diminuant le degré de généralité des questions (p.ex. en éviteant de poser une question sur le trajet, mais deux questions successives sur la gare de départ et la gare d'arrivée à l'instar de ce qui est fait dans [YOUNG 89]).

Troisièmement, la complexité prévisible d'une intervention de l'utilisateur varie avec la complexité de reconnaissance de chacun de ces actes, pris isolément. Par exemple, lors d'un choix proposé à l'utilisateur, il est intéressant de diminuer le nombre de réponses strictes entre lesquelles il a à faire son choix. Une réponse est considérée comme stricte si elle répond à la question et n'ajoute pas d'information supplémentaire utilisable par le système. Dans un menu, ce nombre est égal au nombre de choix possibles; dans une question ouverte sur un nom de gare, il est égal à cinq mille; dans une question du type "Allez vous dans une de ces trois villes? si oui laquelle : Paris, Marseille ou Lyon ?" il est égal à quatre (Notons que l'utilisateur peut répondre "Non" à cette question). Cela peut aussi être réalisé par le choix d'un codage particulier adapté à l'acquisition d'une information (épellation pour un nom, saisie DTMF sur le clavier du téléphone pour un nombre,...).

Enfin, cette complexité dépend de celle de la formulation de chaque réponse stricte. La principale stratégie consiste à tirer parti du comportement induit chez l'utilisateur par les interventions du système, de manière à provoquer un type particulier de formulation. De façon très simplifiée, il faut produire des questions longues pour obtenir des réponses courtes et produire des questions courtes pour obtenir des réponses longues [MOREL 89]. La mise en oeuvre de cette stratégie nécessite une étude fine des réactions de l'utilisateur en fonction des énoncés choisis pour la tâche qui nous concerne. Dans la même idée mais de façon plus réaliste, un vocabulaire simple permettant de désigner les principaux concepts est introduit (p.ex. l'association d'un numéro à chaque horaire présenté par le système, de manière à pouvoir y revenir au cours du dialogue par simple désignation de ce numéro). Ce vocabulaire comprend aussi les références aux principales fonctionnalités (p.ex. revoir ou modifier une information). Il doit servir de base à l'élaboration d'un langage commun facilement assimilable par l'utilisateur et lui fournir ainsi les moyens d'exprimer ses besoins de façon concise.

2.1.3 Expression de la prise d'initiative

Pour qu'une prise d'initiative issue de l'utilisateur soit prise en compte, elle doit être exprimée à travers une intervention reconnaissable, ce qui est généralement incompatible avec l'explosion de l'espace de recherche qu'elle implique.Cette prise d'initiative peut avoir plusieurs origines : l'intention de demander une pause dans le dialogue pour parler à quelqu'un d'autre, l'intention de préciser l'action suivante à réaliser de façon plus complète que ce qui est demandé par le système, l'intention d'interrompre une action en cours pour en réaliser une nouvelle (dans le cadre d'un déroulement sans erreur du dialogue), d'entamer un processus de vérification voire de correction des informations déjà "validées"...
Nous allons voir comment diminuer la complexité prévisible des interventions de l'utilisateur dans ces deux derniers cas de prise d'initiative. Pour ce faire, nous introduisons la notion de phase.

2.1.3.1 Notion de phase

Du point de vue de l'utilisateur, une phase représente un état du dialogue. A chaque phase est

associé un ensemble d'actes. L'utilisateur sait, au moyen d'une première annonce lors de l'entrée dans la phase et par un accès éventuel au guide, qu'il est autorisé à réaliser toute action de cet ensemble tant qu'il est dans cette phase. Tout changement de phase lui est signalé. Le principal intérêt de cette notion réside dans la simplification de la représentation mentale que peut avoir l'utilisateur du système. Elle permet de diviser le dialogue en unités distinctes et facilement identifiables, dévolue chacune à un traitement particulier (acquisition de la demande, examen des horaires,...).A l'intérieur d'une phase, le dialogue apparaît relativement libre. L'utilisateur peut, notamment, utiliser les possibilités de sur-information lors de l'acquisition de différents concepts.

2.1.3.2 Contrôle de l'enchaînement des actions

Pour que la prise de contrôle de l'enchaînement des actions exercée par l'utilisateur soit efficace, elle doit être soutenue par une représentation simple des enchaînements possibles entre ces différentes actions. La solution adoptée consiste à utiliser une structure arborescente de phases qui constitue la base du modèle de tâche [SADEK 90]. Cette structure est associée à un ensemble de mots-clés et expressions-clés permettant à l'utilisateur de prendre l'initiative de déplacements dans cette structure comme p.ex. *"Retour"* ou *"Sommaire"*.

Le passage entre les différentes phases peut être direct (descente d'un niveau dans l'arborescence, utilisation d'une expression-clé servant de raccourci pour passer d'une branche à l'autre de l'arborescence) ou indirect en cheminant dans l'arborescence par le mot-clé *"Retour"*. Une même phase peut se trouver dupliquée à plusieurs niveaux dans l'arborescence faisant ainsi fonction d'appel de sous-tâche. Le mot-clé *"Retour"* est associé à un parcours de l'arborescence dirigé des feuilles vers la racine. Les fonctionnalités attachées à ce mot sont la terminaison normale d'une tâche ou son abandon. La gestion en cas d'abandon des informations recueillies dans cette phase n'est pas du ressort de l'utilisateur (sauf procédure explicite engagée par le système). Un cas particulier se présente si la phase en cours ne peut être abandonnée ou terminée sur initiative de l'usager. Un exemple présent dans notre application est l'entrée en phase *"épellation"*. Celle-ci a un caractère obligatoire. La phase l'ayant introduite (phase *"acquisition"* par exemple) ne peut être poursuivie tant qu'elle ne s'est pas terminée totalement. Pour l'utilisateur, tout se passe comme si la phase *"épellation"* avait temporairement remplacé la phase *"acquisition"*. Le mot-clé *"Retour"* ne doit donc pas renvoyer à la phase *"acquisition"* mais à la deuxième phase rencontrée dans l'arborescence ne présentant pas ce caractère obligatoire.

Cette division en phases distinctes à pour but d'empêcher l'usage de sur-information d'une phase à l'autre hormis celle autorisée. En effet certains changements de phase font exception à cette règle de non transmission de sur-information d'une phase à l'autre. C'est le cas, par exemple dans l'intervention suivante :"je voudrais modifier mon heure de départ" qui, si elle est prononcée dans la phase *"examen des horaires"* implique l'activation de la phase *"acquisition/modification de la demande"*avec en sur-information le concept *"heure de départ"*.

Cependant, cette division en phases entraîne une certaine perte d'efficacité, lorsque les interventions que souhaiterait émettre l'utilisateur nécessitent un changement de phase préalable à leur émission, par l'emploi d'au moins un mot-clé. De plus, il n'est pas toujours possible de mettre cette division en évidence. Par exemple, une division supplémentaire de la phase *"acquisition de la demande"*, regroupant l'acquisition des cinq principaux paramètres d'une requête d'horaires (gares de départ et d'arrivée, date et heure approximative de départ ou d'arrivée) semblerait difficile à mettre en oeuvre si les performances de reconnaissances l'exigeait, sans passer dans un dialogue complètement dirigé. L'usage de raccourcis risque de rendre difficile à appréhender la structuration arborescente sous-jacente.

Le mot-clé *"Retour"* est associé à un parcours à l'intérieur de l'arborescence des feuilles vers la racine. Il ne permet en aucun cas de parcourir le chemin inverse dès lors que le chemin emprunté est un raccourci ou une remontée vers la racine. Cette action est cependant nécessaire pour rectifier une erreur lors d'un de ces déplacements. Ce problème est directement lié au contrôle de la validité des informations interprétées par le système.

2.1.3.3 Le contrôle des informations

L'intérêt de la prise d'initiative par l'utilisateur prend toute son importance dans l'usage qu'il peut en faire afin d'entamer lui-même un processus de vérification et de correction éventuelle des informations qui ont été transmises au serveur. Ceci

peut se réaliser de manière spontanée (comme p.ex. "je ne vais pas à Paris, je vais à Rennes"). Cette spontanéité ne peut plus être autorisée dès lors qu'un changement de phase se produit entre-temps. La mise en oeuvre de ce processus nécessite donc de pouvoir interrompre le dialogue, revenir éventuellement à la phase précédente, désigner des concepts déjà évoqués au cours du dialogue (gare, heure de départ,...) et désigner des opérations à effectuer sur ces concepts (revoir ou modifier). Pour des raisons d'efficacité la reprise de la dernière intervention réalisée par l'usager semble devoir être offerte de manière systématique. La meilleure solution est l'emploi d'un mot-clé comme *"Annulation"*. Celui-ci offre entre autre la précieuse possibilité d'annuler un parcours intempestif ascendant ou transversal dans l'arborescence. Cependant cette solution, employée systématiquement, entraîne une sous utilisation des capacités d'expression liées à l'usage de la parole continue. La spécification de la phase *"examen des horaires"* permet de mettre en évidence ce problème. En effet, au moment de la présentation des horaires à l'utilisateur, il faut pouvoir proposer le prix du trajet. Ce prix peut être calculé en fonction d'un taux de réduction et de la classe choisie pour le voyage. L'utilisateur doit avoir aussi la possibilité de demander le détail de l'horaire qu'on lui présente dans un premier temps sous forme simplifiée. Il peut aussi demander à consulter l'horaire suivant ou le précédent. L'enchaînement de ces diverses actions est a priori non contraint. Pour des raisons d'efficacité, il est intéressant de regrouper ce qui concerne la demande de prix avec les autres demandes alors qu'elle pourrait faire l'objet d'une phase particulière du fait des sous dialogues d'acquisition des différents paramètres qu'elle est susceptible d'entraîner. La possibilité offerte à l'utilisateur de préciser, de façon sur-informative, la classe et le taux de réduction choisis au moment où il demande son prix, respecte mieux la division en phases, si cette demande reste à l'intérieur d'une phase unique : *"examen des horaires"*. La rectification d'erreurs peut ici se passer du mot-clé *'Annulation'* pour revêtir une forme plus naturelle telle que: "Non je voudrais le détail". Ceci permet d'exploiter au mieux les capacités actuelles du système de reconnaissance.
Certains concepts restent difficile à désigner pour en demander la rectification (comme p.ex. le fait de disposer d'un clavier à touche). Ceci sous-entend de mettre en oeuvre un processus plus général de désignation des concepts. Le parcours d'une liste des concepts acquis, accessible par un mot-clé: *"Vérification"* en est un exemple. Cette solution n'a pas été retenue. L'expérimentation devra permettre d'évaluer le choix qui a été fait.

2.1.3.4 Les problèmes de mise en oeuvre

L'un des principaux problèmes liés à la mise en oeuvre de ces solutions réside dans la création d'un langage artificiel constitué de mots-clés ou d'expressions-clés, que l'utilisateur a à apprendre pour maîtriser le système. Ce langage doit donc être simplifié au maximum. Dans cette solution, il est composé principalement de trois mots-clés : *"Sommaire"* (Retour à la racine de l'arborescence) *"Retour"* (déplacement vers la racine) et *"Annulation"* (annulation pour reprise de la dernière intervention de l'utilisateur). A cela s'ajoutent quelques expressions clé : *"je voudrais modifier" "je voudrais revoir"*. La question se pose de savoir si l'utilisation conjointe de deux mots-clés *"Retour"* et *"Annulation'* dont les effets sont parfois très similaires est raisonnable et ne risque pas d'entraîner des confusions préjudiciables à la poursuite du dialogue. Une solution consiste à n'utiliser effectivement qu'un des deux mots-clés et à poser une question pour connaître l'intention exacte de l'utilisateur telle que "Voulez vous revenir à la rubrique horaire ou annuler votre dernière intervention?". L'expérimentation devra aussi permettre d'élucider ce point.

2.2 Détection et correction

Hormis le processus de vérification et de correction dont l'utilisateur peut prendre l'initiative, le système doit posséder ses propres moyens pour détecter et corriger les erreurs. Cette détection d'erreurs s'appuie sur la vérification de contraintes de cohérence du dialogue et la gestion des confirmations. Le processus de correction d'erreurs variera ensuite selon une estimation de la confiance qu'il peut avoir dans les interventions que le système a reconnu.
Les contraintes de cohérence codent trois types d'information: (1) certains paramètres ne peuvent avoir qu'une valeur; (2) une valeur infirmée pour un paramètre ne peut être acceptée de nouveau pour ce même paramètre que sous certaines conditions; (3) les valeurs de plusieurs paramètres peuvent être dépendantes les unes des autres.

La gestion des confirmations envisagée comporte trois volets. Le premier volet concerne la confirmation systématique pour les paramètres principaux de la requête d'horaires ainsi que pour la demande de prix. Cette confirmation peut prendre plusieurs formes. Elle peut faire l'objet d'une question directe ou être implicitée dans la question

suivante du système. Le choix est effectué en fonction du nombre d'erreurs déjà constaté. Le deuxième volet concerne la mise en place d'une possibilité de vérification de l'ensemble des paramètres de la requête d'horaires proposée à l'utilisateur avant le lancement de la recherche dans la base de données. Le troisième volet concerne les interventions qu'on juge inopportun de faire confirmer, afin d'accélérer le dialogue. La gestion des erreurs est dans ce cas laissée à l'utilisateur qui se chargera éventuellement de procéder à une annulation. C'est le cas pour les déplacement dans l'arborescence. Ceci est justifié, entre autres, par le fait qu'il est difficile de gérer un dialogue de correction d'erreurs en ces points particuliers.

Les interventions reconnues peuvent ne pas être traitées de la même façon lorsqu'elles induisent des incohérences. Ainsi, une valeur confirmée peut ne pas être remise en cause immédiatement si une nouvelle valeur apparaît de manière sur-informative, dans une réponse à une autre question. En revanche elle sera immédiatement prise en compte si elle apparaît sous la forme d'une infirmation telle que "je ne ... pas ...".
En cas d'erreur persistante, le système met en oeuvre les techniques de prévention des erreurs vues auparavant (épellation,...) afin de diminuer la complexité des interventions de l'utilisateur .

3 L'ETAPE DE VALIDATION

La validation des choix proposés dans ce papier doit s'effectuer en plusieurs étapes. Une première validation va être réalisée à travers la constitution d'une maquette mono-locuteur. Elle va servir à observer les comportements du système et noter les premières insuffisances. Ensuite, une expérimentation en auprès du "grand public", mais dans un cadre restreint (renvoi des utilisateurs vers un comparse en cas d'échec) sur environ 200 personnes, avec reproduction des erreurs relevées précédemment, aura lieu. Cette validation va servir à valider les différentes solutions, mais aussi à recueillir un corpus comportant des exemples de prises d'initiative de la part d'un usager, face à un système informatique, ainsi que des exemples des processus employés par les utilisateurs pour corriger les erreurs survenant dans le dialogue.

4 CONCLUSION

Les principales conséquences de la mise en oeuvre des principes évoqués ci-dessus sont une réduction de la liberté consentie à l'utilisateur par rapport à un usage naturel de la parole. Le langage artificiel "imposé" à l'utilisateur a été conçu sur la base d'un compromis entre les limites acceptables de liberté d'expression et les contraintes d'efficacité et de robustesse auxquelles le système doit répondre. Les solutions proposées ici se sont révélées particulièrement adaptées à la mise en oeuvre de ce service précis. La structure du modèle de tâche ainsi dégagé permet d'envisager une extension maîtrisable du point de vue de la reconnaissance. L'étape de validation permettra d'effectuer les choix préalables nécessaires à une réalisation multi-locuteur.

BIBLIOGRAPHIE

[GUYOMARD 91] Guyomard M., Siroux J. Cozannet A.
The role of dialogue in speech recognition the case of the yellow pages system In: *Proceedings of Eurospeech-91*, Gènes,Italia,1991.

[JOUVET 91] Jouvet D., Barktova K., Monné J. On the modelization of allophones in HMM based speech recognition system, In: *Proceedings of Eurospeech-91*, Gènes,Italia,1991.

[MOREL 89] Morel M.A.
Computer human-communication In *The structure of multimodal dialogue*,Ed by Taylor M.M., Néel F. & Bouwhuis D.G. eds ,North-Holland, 1989

[SADEK 90] Sadek M.D.
Logical task modelling for Man-machine dialogue. In *Proceedings of the eighth AAAI Conference, Boston*, MA, 1990.

[SADEK 91] Sadek M.D.
Dialogue acts are rational plans. In Proceedings of Venaco II workshop on *"The Structure of multimodal dialogue"*,Mareta,Italy,1991.

[TOULAROHAT 92] Toularohat M.
à paraître dans *Voice system Worldwide '92* Hanoover, 1992

[YOUNG 89] S.J. Young, C.E. Proctor.
The design and implementation of control in voice operated database inquiry systems,In *Computer Speech and Language*, 1989 3, 329-353

19es J.E.P. - Bruxelles, 19 au 22 mai 1992.

UN MODÈLE DE REPRISE DES ERREURS POUR LE DIALOGUE ORAL HOMME-MACHINE

PIERRE NERZIC - nerzic@merlin.enssat.fr

IRISA/LLI - ENSSAT - BP 447 - LANNION

Résumé

Cet article décrit un modèle pour la détection et la réparation des erreurs pouvant survenir dans les dialogues oraux homme-machine. Ce modèle est fondé sur le paradigme de la reconnaissance de plans d'actes de langage. Il est composé de trois parties. La première permet la reconnaissance de plans invalides grâce à une bibliothèque de métaplans. Un plan attribué par la machine à l'utilisateur est invalide si ce plan ne respecte pas les règles de construction et d'exécution de plan. La bibliothèque de métaplans nous permet de représenter une grande variété de plans invalides. La seconde partie est une extension de la première partie conçue pour manipuler les références employées dans les dialogues. Notre modèle permet actuellement de prendre en compte les descriptions définies. La troisième partie de notre modèle est une procédure de correction des erreurs détectées dans les plans et de traitement des références ambiguës incorrectement résolues. Ces corrections sont rendues possibles par la mémorisation de points de retour arrière.

INTRODUCTION

La parole autorise une très grande liberté d'expression dans les dialogues homme-machine mais en contrepartie entraîne un plus grand risque de mauvaise compréhension entre les interlocuteurs à cause des impératifs de convivialité à offrir à l'utilisateur.

Deux problèmes caractéristiques des dialogues oraux ne sont pas traités par les modèles actuels. La structure des modèles de dialogue actuels est à l'origine du premier qui concerne la reconnaissance et la manipulation de plans utilisateur erronés (précondition fausse, ordonnancement des actions invalide, etc.).

Dans l'exemple de la figure 1, U constate que l'une de ses actions physiques est en échec, il demande à S de l'aider. Ce dernier s'aperçoit alors que dans le plan du domaine *enlever(U,roue)* attribué à U, il manque une sous-action pour démonter les boulons. Nous dirons alors que du point de vue de S, le plan de U est invalide.

U se fait aider par S pour démonter une roue de voiture :
U: « Je tire sur la roue mais elle ne vient pas »
S: « Il faut d'abord enlever les boulons! »

Figure 1 : échec dans un dialogue

Plus généralement, nous appelons plan invalide tout plan qui ne respecte pas les règles de construction et d'exécution de plans utilisées par la machine. Ces règles interdisent par exemple l'exécution d'une action dont les préconditions sont fausses. Nous appelons action invalide, toute action qui n'aurait pas dû être faite par l'utilisateur compte tenu des intentions qui lui ont été attribuées ou qui est mal faite du point de vue machine.

Cet article propose une méthode qui permet de résoudre le problème de l'expression des altérations d'un plan valide d'une manière générique, complètement indépendante du domaine considéré. Cette méthode est fondée sur le concept des métaplans déjà largement utilisé par Litman [LIT85] pour réaliser notamment le suivi des dialogues.

Le second problème des dialogues homme-machine concerne certaines références (anaphores, descriptions définies) qui peuvent être mal comprises ; leur identification fait appel à des mécanismes souples qui peuvent, en se trompant, entraîner des malentendus dont la prise de conscience par les interlocuteurs peut être tardive. Cet article présente un modèle permettant de tenir compte d'éventuels malentendus entre les interlocuteurs. Si dans l'exemple précédent, U pense à la roue de secours, qui est fixée par une barre, la réponse du système va lui parait incompréhensible. Notre modèle permet de placer des points de retour en arrière dans les plans du dialogue permettant ainsi de réexaminer certaines références incorrectement résolues sous un éclairage différent.

Enfin, cet article présente un module permettant de réparer les plans invalides et de réparer les conséquences d'une mauvaise résolution des références.

Cet article est composé de quatre parties. La première décrit le modèle permettant la reconnaissance de plans invalides. La seconde partie présente l'extension permettant de manipuler les références employées dans les dialogues. La troisième partie décrit la procédure de correction des erreurs détectées dans les plans. Enfin, la dernière partie explique le traitement des références incorrectement résolues.

RECONNAISSANCE DES INTENTIONS

PRÉSENTATION DU MODÈLE

Pour décrire les nombreuses altérations qui peuvent affecter les plans de l'utilisateur, nous utilisons un cadre indépendant de toute application, constitué d'une bibliothèque de métaplans et d'un algorithme de reconnaissance de plans. Ce couple travaille sur des séquences d'observations relevées par le système et délivre un diagnostic analysant les éventuels problèmes présents dans ces séquences. Ce diagnostic est ensuite utilisé pour corriger spécifiquement les erreurs détectées.

Nous avons recours aux métaplans car c'est un moyen simple pour représenter de nombreuses propriétés des plans des utilisateurs. Les métaplans sont constitués comme les plans normaux mais leurs préconditions, contraintes, décompositions et effets manipulent celles et ceux des plans normaux par des opérateurs particuliers.

La bibliothèque de métaplans de notre modèle permet la représentation de deux procédures suivies par l'agent pour exécuter un plan. La première de ces procédures consiste à vérifier toutes les préconditions du plan, la seconde consiste à exécuter toutes les sous-actions de la décomposition du plan.

Ces deux procédures qui sont implicites dans les modèles classiques, sont représentées explicitement dans notre modèle par deux classes de métaplans : *exécute(Agent, Plan, Diagnostic)* et *vérifie(Agent, Proposition, Diagnostic)*, chaque classe possédant plusieurs variantes. Ces variantes consistent à formuler plusieurs cas particuliers pour ces métaplans ayant chacun un ensemble de contraintes portant notamment sur l'ordre d'exécution des sous-actions et la valeur de vérité des préconditions. Chaque cas particulier fait l'objet d'un métaplan dans la bibliothèque.

L'entête des métaplans contient trois paramètres :

— le nom de l'agent qui réalise ce métaplan ;

— le plan du domaine de l'agent qui est réalisé à travers le métaplan ou le prédicat qui est vérifié par l'agent ;

— un diagnostic qui décrit l'état d'avancement du plan de l'agent, par exemple : *début* signifie que l'agent en est au début de l'exécution de son plan.

Ce formalisme permet, en jouant sur l'expression des contraintes, de décrire des plans invalides. Ce point est développé dans la seconde partie de ce chapitre.

La figure 2 donne la liste de quelques métaplans qui permettent de reconnaître des plans valides. Voici une description des prédicats utilisés dans les contraintes de ces métaplans :

— *précond_de(Prec, Action)* est vrai si *Prec* est une des préconditions de *Action* ;

— *action_de(Prop, Action)* est vrai si *Prop* est un effet de *Action* ;

— *effet_de(Action1, Action2)* est vrai si *Action1* est dans le corps de *Action2* ;

— *avant_dans(Action1, Action2, Action)* est vrai si *Action1* est avant *Action2* dans le corps de *Action* ;

— *tous(Pred1, Pred2)* est vrai si toutes les instanciations des variables qui rendent *Pred1* vrai, permettent de vérifier *Pred2* ;

— *non(Pred)* est vrai si *Pred* n'est pas vrai (négation par l'échec à la manière de Prolog) ;

— *faite(Action)* est vrai si l'observateur du plan à reconnaître (la machine) a déjà observé *Action*.

M2 entête: exécute(Agent, Plan, début)
corps : exécute(Agent, Action, Diagnostic)
contraintes :
action_de(Action, Plan)
non(avant_dans(Action1, Action, Plan))
tous(précond_de(Prec,Plan), vrai(Prec))

M3 entête: exécute(Agent, Plan, en_cours)
non-détaillé
M4 entête: exécute(Agent, Plan, fini)
corps : exécute(Agent, Action, Diagnostic)
contraintes :
action_de(Action, Plan)
faite(Action1)
avant_dans(Action1,Action, Plan)
non(avant_dans(Action, Action2, Plan))
tous(précond_de(Prec,Plan), vrai(Prec))

Figure 2 : sous-ensemble de la bibliothèque

Le métaplan M2 décrit l'exécution de la première action de *Plan*. La première contrainte indique que *Action* doit faire partie de ce plan, la seconde précise que *Action* ne doit pas être précédée par une autre action et la troisième indique que toutes les préconditions de *Plan* doivent être vraies. Les autres métaplans sont similaires à celui-ci.

Avec cette bibliothèque de métaplans, nous utilisons une méthode abductive de reconnaissance de plans fonctionnant d'une manière très classique. Notre modèle possède des capacités de reconnaissance incrémentale grâce aux contraintes qui concernent les actions exécutées auparavant : contrainte *faite(Action)*.

RECONNAISSANCE DES PLANS INVALIDES

Le modèle que nous proposons permet, outre la représentation et la reconnaissance de plans valides, une expression très simple d'un certain nombre de défauts qui peuvent affecter les plans qui sont attribués à l'utilisateur par la machine.

Notre modèle autorise la représentation d'erreurs concernant notamment :

— l'ordonnancement des actions dans un plan : action manquante, action en trop, actions désordonnées ;

— l'existence de préconditions fausses ou superflues.

Chaque type d'erreur s'exprime par un ou plusieurs métaplans. La figure 3 en montre deux exemples : le

métaplan M7 détecte les actions interdites à cause d'une précondition fausse et le métaplan M9 reconnait les tentatives infructueuses pour satisfaire une proposition.

M7 entête : exécute(Agent, Plan, faux(Préc))
corps : exécute(Agent, Action, Diagnostic)
contraintes :
action_de(Action, Plan)
précondition_de(Préc, Plan)
faux(Préc)

M9 entête : vérifie(Agent, Prop, faux(Prop))
corps : exécute(Agent, Plan, Diagnostic)
contraintes :
effet_de(Prop, Plan)
faux(Prop)

Figure 3 : quelques métaplans de détection d'erreur

Les figures 4a et 4b développent un exemple significatif traitant de phrases issues d'un dialogue réaliste avec actions physiques imbriquées. Dans cet exemple, l'utilisateur U décrit à la machine les actions physiques qu'il ne parvient pas à faire et la machine essaie de trouver ce qui empêche ces actions physiques. Les actions physiques réalisées par U sont inférées directement du contenu propositionnel des actes de langage.

entête : démonte(Agent, Roue)
préconditions : libre(Roue)
corps : enlève(Agent, Roue)

entête : libère(Agent, Roue)
corps : retire(Agent, Boulons)
contrainte : boulons_de(Boulons, Roue)
effet : libre(Roue)

entête : retire(Agent, Boulons)
corps : dévisse(Agent, Boulons)
pose_terre(Agent, Boulons)

a) actions du domaine

observation de :
U: « Je n'arrive pas à enlever la roue ... »
(sans autres actions auparavant)
elle est formalisée par le métaplan suivant :
exécute(U, enlève(U,roue1), fini)
avec sait(S, faux(libre(roue1)))

• exécute(U, démonte(U, roue1),
faux(libre(roue1))) par M7
• exécute(U, enlève(U,roue1), fini)

b) détection d'une erreur

Figure 4 : Détection d'une erreur dans les plans de l'agent

La figure 4b montre la reconnaissance d'un plan invalide. L'algorithme de reconnaissance abductive tente d'expliquer l'observation en la recherchant dans le corps d'un des métaplans dont les contraintes sont vérifiées, ici m7 convient et signifie : l'action de poser la roue par terre est prématurée car la précondition *libre(roue1)* est fausse.

RÉSOLUTION DES RÉFÉRENCES

La partie précédente vient de décrire la manière dont notre modèle reconnaît une grande variété de plans invalides sous-jacents aux énoncés de l'utilisateur dans un dialogue. Nous allons maintenant examiner certains problèmes que posent l'emploi de références par les interlocuteurs et comment notre modèle permet d'y apporter une solution.

Le principal problème à résoudre est posé par l'ambiguïté des références employées dans un dialogue. La construction et la résolution de ces références (descriptions définies, anaphore, etc.) nécessitent chez chaque interlocuteur, un contexte attentionnel [GS86] qui contient l'ensemble des entités présentes consciemment dans l'esprit de l'interlocuteur, comme les boulons, la roue ou même la voiture des exemples précédents. Ce contexte permet au locuteur de construire des références concises et à l'auditeur de choisir l'entité réellement désignée par le locuteur. Le problème se ramène donc à déterminer les contextes attentionnels des interlocuteurs ; le contexte de l'auditeur devant se rapprocher autant que possible de celui du locuteur.

DESCRIPTIONS DÉFINIES

Nous nous sommes pour l'instant limités au cas particulier des descriptions définies ce qui nous permet de simplifier l'expression du contexte attentionnel et nous traitons les problèmes d'ambiguïté dus à une définition plus ou moins complète de ces descriptions. Nous avons choisi une représentation de ces références par un formalisme permettant de les traiter directement en Prolog. Ce formalisme est un sous-ensemble de celui des formes quasi-logiques utilisé par Alshawi [ALS90].

le(Référent, Type, Liste)

« le boulon de la roue arrière droite de la voiture »
le(Boulon, boulon, [
le(Roue, roue, [
le(Voiture, voiture, [ ]),
roue_de(Roue, Voiture),
position_de(arrière_droit, Roue)]),
boulon_de(Boulon,Roue)])

Figure 5 : description définie

Ce prédicat possède trois paramètres, le premier symbolise le référent désigné par la description (libre ou instancié), le second est un symbole donnant le type du référent et le dernier paramètre est une liste décrivant certains attributs du référent. Cette liste peut être réduite

à la liste vide, comme dans le cas de la voiture de l'exemple.

La résolution de ce prédicat est faite grâce à un examen du contexte attentionnel de l'auditeur et à des mécanismes souples comme l'inférence par défaut ou des heuristiques. Cette résolution fournit aucun, un ou plusieurs symboles pour la variable Référent. Nous expliquons un peu plus loin ce qui est fait lorsque la résolution ne fournit pas exactement un seul résultat.

Dans le cas des modèles classiques, comme celui de Litman et Allen [LA84], la résolution des références est faite avant de procéder à la reconnaissance des intentions et les références présentes dans les observations sont définitivement remplacées par le référent qui a été trouvé à ce moment là. Deux problèmes apparaissent avec une telle résolution préalable. Le premier concerne le choix du contexte dans lequel s'effectue la résolution puisqu'il s'effectue arbitrairement : le contexte choisi est celui du plan actif [CAR88]. Cela empêche donc de pouvoir reconnaître un référent dans le cadre d'un plan invalide.

Le second problème est rencontré fréquemment dans les dialogues oraux lorsqu'on désire remettre en cause le résultat de la résolution d'une des références. Comme les références sont substituées par le référent qui en a été obtenu à un moment précis (par exemple, *le(Boulon,...)* est remplacé par *boulon1* lorsqu'il était question de l'une des roues), leur mode de désignation est perdu (pronom, description définie, etc.) et ces références ne peuvent plus faire l'objet d'une éventuelle résolution dans un autre contexte.

Pour un traitement correct du second problème, nous conservons dans les actions la forme instanciée des descriptions définies. De cette manière, elles sont distinguées des simples constantes et peuvent faire l'objet d'une autre résolution avec un contexte différent.

Une solution au premier problème est décrite dans les paragraphes suivants.

CONTEXTE ATTENTIONNEL

Nous proposons, dans le cadre de notre modèle, une solution permettant à la fois de représenter le contexte attentionnel servant à la résolution des descriptions et permettant de retraiter les références employées par les interlocuteurs. Cette solution consiste à placer dans l'entête des plans et métaplans des informations supplémentaires destinées à représenter le contexte attentionnel présent à un niveau donné. Ces informations se répartissent en deux catégories.

D'un coté, on trouve les informations sûres à propos du contexte : ce sont les éléments qui sont présents dans les descriptions définies comme des constantes (valeurs numériques, noms propres, etc.) ou les types des objets. Si une autre résolution devait avoir lieu, ces éléments seraient utilisés tels quels.

De l'autre coté, on trouve les informations qui ne sont pas certaines, ce sont celles qui ont été déduites ou supposées par le module de résolution de références (valeurs par défaut, etc.). L'exemple précédent contient une telle information : la voiture n'est pas définie et l'algorithme de résolution est amené à supposer que la voiture en question est celle de l'agent. Nous appelons ces dernières informations : hypothèses sur le contexte attentionnel. Elles sont formulées par l'algorithme de résolution des références lorsque plusieurs référents correspondent à une même description. Dans ce cas, cet algorithme a deux possibilités pour désambigüer la référence. La première consiste à rechercher un critère de discrimination entre ces référents en fonction du contexte et des règles mentionnées précédemment (inférences par défaut, heuristiques, etc.). L'algorithme de résolution renvoie à la fois le référent restant et les critères choisis. Dans l'exemple, le critère est fourni par une inférence par défaut : tout objet appartient au locuteur sauf si la mention du contraire est faite.

La seconde possibilité consiste à négocier le référent avec l'utilisateur en obtenant plus d'informations selon la méthode qu'a proposée P. Heeman [HEE91].

REPRÉSENTATION DU CONTEXTE

La résolution d'une référence permet donc d'obtenir deux ensembles d'informations qui permettent de compléter le contexte attentionnel. Nous allons maintenant examiner comment la totalité du contexte est gérée. Grosz et Sidner ont montré que ce contexte évolue en fonction de l'avancement du dialogue [GS86]. A chaque action du plan que suit l'utilisateur correspond un sous-ensemble du contexte attentionnel : les informations apportées par cette action grâce à ses références. L'ensemble du contexte est représentable par une pile dont le sommet contient les éléments concernant l'action la plus spécifique et dont la base contient les éléments concernant l'action la plus générale du plan.

Notre modèle permet de gérer cette pile de la manière suivante. A un point quelconque du dialogue, un plan est en cours, une partie de ses actions est effectuée, une autre partie reste à faire : A_1 suivie de A_2 ; avant A_1, un certain contexte C_E est présent (ancien sommet de la pile), l'observation de A_1 va modifier ce contexte : les informations obtenues dans les références présentes dans cette action : C_1 s'ajoutent à C_E et donnent le contexte C'_1 dans lequel vont être reconnues les références de A_2. Les hypothèses qui sont formulées au cours des différentes résolutions de références sont accumulées dans les contextes successifs et ne sont pas dépilées. Cette technique autorise une correction ultérieure des erreurs de résolution car elle permet de localiser l'apparition des hypothèses concernant les références.

L'intégration de ce mécanisme de pile dans notre modèle s'inspire du modèle de D. Litman [LIT85] dans lequel le contexte attentionnel est contenu dans les paramètres des plans. De manière comparable, nous plaçons deux paramètres supplémentaires dans l'entête des plans que nous appelons respectivement contexte entrant et contexte sortant. Le contexte entrant est le contexte de dialogue disponible avant d'effectuer l'action. Le contexte sortant est la réunion des hypothèses présentes dans le contexte entrant et de celles qui ont été faites lors de la résolution des références de l'action. La figure 6 montre un prototype de plan.

plan(..., C_E, C_S)
corps:
plan1(...,C_E, C_1)
plan2(...,C_1,C_2)
plan3(...,C_2,C_S)

plan1(paramètres, C_E, C_S)
contrainte :
C_S = C_E + hypothèses faites pour résoudre les références des paramètres

figure 6 : prototype montrant la gestion du contexte

répare(Agent,Roue,C_E,C_S)
corps: installe_cric(Agent,C_E,C_1)
demonte(Agent,Roue,C_1,C_2)
enlève_cric(Agent,C'_2,C_S)
contrainte: C'_2=C_2 - informations relatives à la roue

installe_cric(Agent,C_E,C'_S)
corps: prend(Agent,Cric,C_E,C_1)
met_en_place(Agent,Cric,C_1,C_S)
contrainte: C'_S=C_S - informations relatives au cric

a) exemple de plans

observation de :
U: « Je prends le cric »
formalisée par : prend(U, le(Cric, cric, [ ]))

résultat final, après reconnaissance et résolution, sans les métaplans :
- répare(U,Roue, [ ],
  [voiture_de(le(voiture1, voiture, [ ]), U)])
  - installe_cric(U, [ ],
    [voiture_de(le(voiture1, voiture, [ ]), U)])
    - prend(U, le(cric1, cric, [ ]), [ ],
      [voiture_de(le(voiture1, voiture, [ ]), U),
      cric_de(le(cric1, cric, [ ]), voiture1)])

b) première action

observation de :
U: « Je mets le cric en place » (à la suite)
formalisée par :
met_en_place(U, le(Cric, cric, [ ]))

résultat final sans les métaplans :
- répare(U,Roue, [ ],
  [voiture_de(le(voiture1, voiture, [ ]), U)])
  - installe_cric(U, [ ],
    [voiture_de(le(voiture1, voiture, [ ]), U)])
    - met_en_place(U,le(cric1, cric, [ ]),
      [voiture_de(le(voiture1, voiture, [ ]), U),
      cric_de(le(cric1, cric, [ ]), voiture1)],
      [voiture_de(le(voiture1, voiture, [ ]), U),
      cric_de(le(cric1, cric, [ ]), voiture1)])

c) deuxième action

figure 7: exemples de gestion du contexte

Chaque plan du domaine présent dans la bibliothèque définit les échanges qui sont effectués entre son contexte entrant et son contexte sortant, par des contraintes. Des raisons d'efficacité conduisent à définir également les échanges qui ont lieu entre chacune de leurs actions. La figure 7a montre de telles spécifications.

La signification des hypothèses des actions de bas-niveau est : la voiture appartient à U et le cric est celui de cette voiture. La spécification des échanges de contexte entre les actions montre que les incertitudes quant au cric ne remontent pas plus haut que l'action le concernant. En revanche, les incertitudes concernant la voiture persistent tout au long du plan car toutes les opérations ultérieures dépendent de cette voiture (démontage de la roue, etc.).

Dans la figure 7b, la référence au cric est résolue, après reconnaissance de plans, dans le contexte entrant du plan qui englobe l'observation, soit [ ]. Cette résolution produit deux hypothèses qui seront fournies au plan qui explique la seconde observation dans la figure 7c.

Les exemples de la figure 7 sont simplifiés et font abstraction des métaplans expliqués dans le chapitre précédent. Dans le cas général, la résolution d'une référence présente dans une action (*exécute* ou *vérifie*) se fait alors en fonction du métaplan dans lequel cette action sera placée. C'est le prédicat *faite(ActionPrécédente)* qui permet de retrouver facilement le contexte de l'action précédente, contexte dans lequel il faut résoudre les références de l'action courante.

CORRECTION DES PLANS INVALIDES

Dans cette partie, nous décrivons les processus de correction des erreurs découvertes par notre modèle.

Lorsqu'une séquence d'action est reconnue invalide, une procédure de réparation est déclenchée. Cette procédure commence par vérifier qu'il n'y a pas de malentendu sur les référents des actions incriminées, puis tente de replanifier les buts de ces actions invalides.

La première phase de cette procédure consiste en une vérification des référents des actions jugées invalides. Un malentendu sur un référent peut par exemple faire croire à la machine qu'une des préconditions d'une action observée est fausse et ainsi entraîner une détection d'erreur non fondée. Notre modèle a été étudié pour faciliter ce travail de vérification des références. Il suffit en effet de vérifier les différentes hypothèses qui ont été proposées au cours de la reconnaissance de plans en se limitant à celles qui sont concernées par le diagnostic de l'erreur.

Lorsqu'une de ces hypothèses sur le contexte se révèle fausse, il nous suffit de corriger la partie du plan dans laquelle cette hypothèse a été formulée. Cette correction consiste à remplacer si c'est possible les référents incorrectement résolus par ceux qui étaient réellement désignés par l'utilisateur. Le chapitre suivant décrit cette correction.

Après avoir vérifié toutes les références suspectes et confirmé le diagnostic, la procédure de correction entame la seconde phase de correction. L'objectif de la machine est de permettre à l'utilisateur de vérifier le plus grand

nombre de ses buts. Pour cela, la machine commence par essayer de planifier le but qui aurait dû être atteint par le plan en échec. Dans le cas où une solution est trouvée, elle est présentée à l'utilisateur. Dans le cas défavorable, la machine essaye de trouver une autre manière de réaliser le plan plus général qui englobe celui qui est en échec.

Le remplacement d'un plan par un autre s'effectue dans le cadre d'une négociation avec l'utilisateur, grâce à des plans d'actes de langages inspirés par ceux de Litman (comme *correct-plan* dans [LIT85]).

RÉFÉRENCES INCORRECTEMENT RÉSOLUES

Cette partie décrit la manière dont notre modèle permet la remise en question des hypothèses faites sur le contexte attentionnel du dialogue. Cette remise en question est entraînée par deux cas de figure.

Le premier cas est rencontré pendant la reconnaissance de plans, lorsque la résolution d'une référence est impossible, dans le cas d'hypothèses trop fortes sur le référent à trouver. Le second cas a été décrit dans le chapitre précédent et apparaît après la reconnaissance de plans, lorsqu'un plan est reconnu comme invalide.

La remise en cause de certaines hypothèses faites sur le contexte du dialogue est faite par le relâchement des hypothèses qui contraignent les références. C'est une technique qui a été décrite par Goodman [GOO87]. Après avoir trouvé un référent et des hypothèses satisfaisantes, il est nécessaire de reconsidérer toutes les autres références qui dépendent de ces hypothèses.

Cet examen reste possible longtemps après la première formulation des hypothèses car les références sont conservées dans leur forme initiale dans les plans. Comme les anciens référents de certaines de ces références sont ainsi remplacés par des nouveaux référents, notre modèle doit aussi revoir les plans qui contiennent ces nouveaux référents. Certains de ces plans concernent des actions de l'utilisateur qui ont été interprétées dans un contexte incorrect. La correction de ces plans est faite de la manière décrite dans le chapitre précédent.

Les autres plans sont ceux de la machine, ils doivent être annulés puis refaits. Nous considérons actuellement que toutes les actions linguistiques (en particulier les actions *informer*) peuvent être annulées, c'est à dire que leurs effets qui consistent en une modification des connaissances de l'interlocuteur, peuvent être supprimés. Nous n'apportons pas de réponse dans le cas d'actions physiques non réversibles. La suppression de ces effets est provoquée par une catégorie d'actions qui signalent à l'utilisateur, par un acte de langage particulier, la présence d'une erreur dans une action précédente. Cette catégorie d'actions qui sont des métaplans, appartient à l'ensemble des plans du dialogue mis en évidence par D. Litman [Lit85]. Voici un exemple montrant l'emploi de ce type d'actions : « Excusez-moi, je me suis trompé quand je vous ai dit que ... ».

Un second acte de langage permet ensuite de mettre en valeur l'action correcte, par exemple : « Je voulais dire en fait que ... ». La suite de la correction consiste à replanifier dans le nouveau contexte les buts de l'utilisateur. Elle se fait par la procédure développée dans le chapitre précédent.

CONCLUSION

Cet article apporte une réponse à deux problèmes importants des dialogues homme-machine. Le premier problème est celui de la reconnaissance intentionnelle de plusieurs catégories de plans invalides. Nous proposons un modèle de représentation et de reconnaissance de plans fondé sur le concept des métaplans. Le second problème est causé par une certaine forme d'ambiguïté du langage naturel, en particulier dans le cas des dialogues oraux. Notre modèle permet une révision de la résolution de certaines références employées par les interlocuteurs. La correction des erreurs détectées se fait grâce à des procédures utilisant les caractéristiques de ce modèle.

RÉFÉRENCES BIBLIOGRAPHIQUES

[ALS90] Alshawi H., Resolving Quasi Logical Forms, in *Computational Linguistics*, Vol 16, Number 3, 1990, 133-144.

[CAR88] Carberry S., Modeling the User's Plans and Goals, in *Computational Linguistics*, Vol 14, Number 3, 1988, 23-37.

[Goo87] Goodman B. A., *Repairing Reference Identification Failures by Relaxation*, Communication Failures in Dialogue and Discourse, R. Reilly editor, North-Holland 87, 123-147.

[GS86] Grosz B. J., Sidner C. L., Attention, Intentions and the structure of Discourse, in Computational Linguistics, Vol 12, Number 3, 1986, 175-204.

[HEE91] Heeman P., Collaborating on Referring Expressions. In *Proceedings of the 29th Annual Meeting of the ACL*, Berkeley, 1991.

[LA84] Litman D. and Allen J., A Plan Recognition Model for Clarification Subdialogues, *Coling84*, Stanford, July 1984, 302-311.

[LC91] Lambert L. and Carberry S., A tripartite plan-based model of dialogue. In *Proceedings of the 29th Annual Meeting of the ACL*, Berkeley, 1991.

[LIT85] Litman D. J., *Plan Recognition and Discourse Analysis: An Integrated Approach for Understanding Dialogues*. PhD thesis, University of Rochester, 1985.

19es J.E.P. - Bruxelles, 19 au 22 mai 1992.

UNE BASE DE DONNEES DE PAROLE HYPERBARE FRANCAISE ET ANGLAISE: PSH/DISPE

ALAIN MARCHAL et CHRISTINE MEUNIER

LABORATOIRE PAROLE ET LANGAGE - URA 261 CNRS
29, av Robert Schuman, 13100 Aix-en-Provence

Résumé

Les mélanges gazeux qui constituent l'atmosphère dans laquelle vivent et évoluent les scaphandriers provoquent un phénomène de dégradation de l'intelligibilité de la parole qui s'amplifie avec la pression. A ce facteur respiratoire viennent s'ajouter les perturbations liées aux équipements de tête des scaphandriers conçus pour la protection contre le milieu ambiant et non adaptés à la parole. Afin de répondre aux exigences de la recherche fondamentale, du développement et de l'évaluation des systèmes de décodage, l'INPP (Institut National de la Plongée Professionnelle) et le CNRS ont entrepris la réalisation d'une base de données de parole hyperbare représentative des tranches de profondeurs opérationnelles. Nous décrivons dans cette communication la composition et la structure de cette base de données dénommée PSH/DISPE.

1-INTRODUCTION.

En plongée profonde, le recours à des mélanges respiratoires gazeux comme l'héliox (hélium + oxygène) permet de s'affranchir des effets narcotiques de l'azote ainsi que des effets convulsifs de l'oxygène. Cependant, l'utilisation de l'hélium, contenu dans ce mélange gazeux associé à l'augmentation de la pression entraîne une forte dégradation de l'intelligibilité de la parole. A ces facteurs respiratoire et hyperbare viennent s'ajouter d'autres éléments qui contribuent à la perturbation de la production de la parole. Les équipements de tête des scaphandriers sont d'abord conçus pour la respiration, mais ils limitent considérablement le mouvement de la mâchoire. Le volume de la cavité du masque provoque divers effets de résonance. La ligne de transmission (filaire en ombilical) est très souvent parasitée. Enfin, le milieu subaquatique et hyperbare constitue un environnement défavorable en raison des différents bruits qui le caractérisent.

Les déformations subies par la parole concernent essentiellement les variations de la fréquence fondamentale (Hollien and Hicks, 1982; Duncan et Jack, 1987, etc.), l'élévation des fréquences de formant des voyelles (Crestel et Guitton, 1987, etc.) et la perte d'énergie des consonnes (Fant et Lindqvist, 1968, etc.). L'ensemble des auteurs ne partage pas toujours les mêmes conclusions quant à l'amplitude des variations de F0 et à la linéarité de l'augmentation des fréquences de formants. Ces divergences peuvent s'expliquer par la grande variété des sources de dégradation de la parole en milieu hyperbare. Les conditions d'enregistrement sont variables (micros, pressions, mélanges gazeux, situations: caisson, tourelle, plongées réelles, plongées simulées, etc.); par ailleurs les corpus de travail, eux-mêmes, sont très différents d'un auteur à l'autre (logatomes, voyelles, mots, phrases, etc.).

Il nous semble que, pour avancer dans la résolution du problème posé par les communication verbales en milieu hyperbare, il faut désormais adopter une approche système; autrement dit, il faut considérer la communication en milieu hyperbare comme une chaîne de communication allant du locuteur jusqu'au récepteur où chaque élément (locuteur et son état, type de micro utilisé, pression ambiante, situation, etc.) est susceptible d'engendrer une

déformation.

C'est dans cette perspective que l'INPP (Institut National de Plongée Professionnelle de Marseille) et l'IPA (Institut de Phonétique d'Aix) ont entrepris la réalisation d'une base de données représentative des conditions de communication en milieu hyperbare et/ou subaquatique.

2-CONSTITUTION DE LA BASE DE DONNEES.

La constitution de cette base de données s'est effectuée en quatre temps:

- Enregistrements en milieu hyperbare à l'INPP et à la Marine Nationale (GISMER, Toulon).
- Sélection des enregistrements les plus représentatifs.
- Segmentation et étiquetage
- Mise au format de base de données.

a) Enregistrements en milieu hyperbare.

Constitution des corpus.

Les corpus ont été conçus afin de permettre:

- la recherche fondamentale et appliquée (modélisation de la parole hyperbare).
- la recherche et le développement de nouveaux produits (fabrication de matériels de type "décodeurs").
- la standardisation de tests d'intelligibilité (évaluation de la qualité de la parole hyperbare).

Ainsi, les corpus de mots anglais (Griffiths, 1967) et français sont élaborés pour obéir au principe de constitution de tests MRT (Modified Rhyme Test, House et al., 1965). Le passage anglais et les phrases françaises, qui contiennent des mots du corpus «mots français», sont phonétiquement équilibrés.

La totalité du corpus est composé de:

- 4 listes de 46 mots français.
- 4 listes de 50 mots anglais.
- 8 phrases françaises.
- un passage anglais.

Enregistrements.

Le corpus a été lu deux fois par 17 locuteurs de sexe masculin (agés de 28 à 38 ans) jusqu'à la profondeur de 300 mètres. Ces locuteurs sont tous des professionnels de la plongée ne présentant pas de pathologie de l'audition. Les enregistrements ont été effectués en premier lieu en condition silencieuse à pression atmosphérique avant la plongée, puis en situation de plongée à différentes profondeurs, et enfin en caisson pendant la phase de décompression. Le mélange gazeux utilisé pour tous les enregistrements effectués en milieu hyperbare était un mélange d'hélium et d'oxygène (l'héliox). Les enregistrements ont été effectués à l'INPP (décembre 1989) ainsi qu'à la Marine Nationale (février à mai 1990) au cours de plongées d'entraînement.

Pour chaque enregistrement sont répertoriés: le numéro de la bande, la date, le code du locuteur, la situation (condition silencieuse/caisson/piscine/tourelle), la profondeur, la pression partielle d'oxygène, le pourcentage d'oxygène, le pourcentage d'hélium, la température du mélange gazeux, la température de l'eau, le type de casque dont le plongeur est éventuellement équipé (KMB-17 ou X-LITE), et enfin le type de micro utilisé (Sennheiser, micro à réduction de bruit et micro Shure en milieu humide).

b) Sélection des enregistrements.

Nous avons retenu pour la base de données les enregistrements réalisés pour les profondeurs de travail les plus fréquentes, soit: 60m, 100m, 180m, 200m, 300m. Afin d'évaluer le problème spécifique de la communication en plongée, des enregistrements réalisés en plongée simulée (piscine) à 180m et à 300m ont également été retenus. Enfin, nous avons retenu cinq locuteurs français reconnus pour parler une variété de la langue exempte de traits régionaux marqués. Il s'agit des locuteurs SO, DC, GL, GD et LY. Pour l'anglais, les deux locuteurs choisis, HS et WN, sont originaires du Hampshire.

c) Segmentation et étiquetage.

L'ensemble des enregistrements sélectionnés a été segmenté manuellement à l'aide de V.E.S.(Espesser et Balfourier, 1989) et étiqueté orthographiquement. Pour les enregistrements bruités (parole à grandes profondeurs), des zones de silence ont été introduites entre chaque mot afin de permettre ultérieurement la mise au format de fichiers de base de données.

d) Mise au format de base de données.

Nous avons choisi le logiciel EUROPEC développé à l'Institut de la Communication Parlée (I.C.P.) à Grenoble (Zeiliger et Sérignat, 1990) afin de réenregistrer nos données. Ce choix a été fait pour deux raisons:

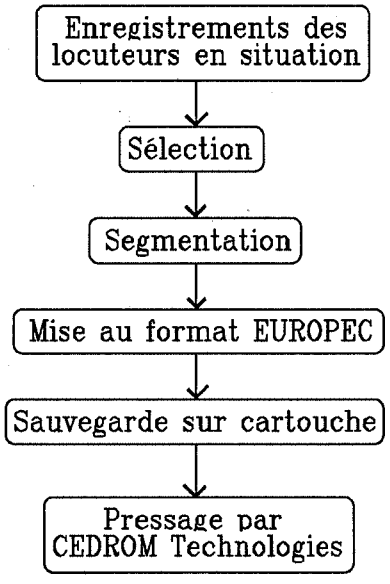
1) Les fichiers issus du logiciel EUROPEC sont conformes au format recommandé par le projet CEE-ESPRIT «SAM» n° 2589 (Fourcin et al., 1989). D'autre part, l'utilisation d'EUROPEC fait de PSH une base de données compatible avec les autres bases de données françaises et

européennes telles que BDSONS (Carré et al., 1987) et EUROM1 (Fourcin et al., 1989). En parallèle de ce logiciel d'enregistrement, l'I.C.P. a développé deux autres logiciels, PTS (Caeroux et Dolmazon, 1990) et GERSONS (Foulard et Sérignat, 1991), en cours d'adaptation à notre base de données. L'environnement issu de cet ensemble de logiciels permet une utilisation très aisée et efficace de la base de données: analyse acoustique et étiquetage (PTS), accessibilité aux fichiers de signal et/ou sélection d'un type de contexte phonétique, d'un ou de plusieurs locuteurs (GERSONS).

Mettre au format EUROPEC les enregistrements déjà éffectués consiste à les «réenregistrer» à l'aide du logiciel EUROPEC. Cette opération se divise en deux étapes. Il est nécessaire tout d'abord de constituer les fichiers d'information (caractéristiques des locuteurs, des corpus, des conditions d'enregistrements, voir tables 1 et 2) qui concernent les données de la base.

Ensuite, viennent les enregistrements proprement dits. Ils consistent en la définition d'un fichier de configuration (fréquence d'échantillonage, type d'entrée, seuils de déclenchement, etc.); puis le logiciel permet d'enregistrer chaque mot (ou séquence) en fonction d'un seuil de déclenchement. A chaque enregistrement (mot ou séquence) est associé une transcription orthographique qui comporte l'adresse du fichier signal. La fréquence d'échantillonage des fichiers d'enregistrements retenue pour notre base est de 40KHz.

Les étapes de la constitution de la base de données peuvent être représentées ainsi:



3-ORGANISATION DE LA BASE DE DONNEES PSH/DISPE.

Stucture des répertoires et sous-répertoires

La base de données est organisée de façon hiérarchique dans l'optique de satisfaire une consultation logique et aisée de la base (voir table 3). A un premier niveau, nous distingons la langue du locuteur. Nous séparons ensuite les enregistrements effectués à pression atmosphérique (à l'air) de ceux réalisés en milieu hyperbare (mélange hélium/oxygène). Deux situations d'enregistrement en milieu hyperbare sont alors distinguées: milieu sec (caisson) et milieu humide (piscine). Enfin, chaque sous-répertoire final représente une profondeur différente.

Dans le répertoire \DOC, on trouve l'ensemble des fichiers qui ont servi à la constitution de la base: les caractéristiques des locuteurs, des corpus, des conditions d'enregistrement (micro, gaz, etc.) et des paramètres acoustiques des enregistrements (fréquence d'échantillonage, seuil de déclenchement, etc.). Le répertoire \DOC contient également un fichier de texte décrivant la base de données.

Structure des noms de fichiers

Certaines informations concernant le contenu des fichiers ne sont pas directement explicites d'après la structure des répertoires; il s'agit du type de corpus prononcé (mots, phrases, passages), du code du locuteur et du type de fichier (signal échantillonné, ou transcription orthographique). Toutefois, ces informations sont présentes dans le nom des fichiers eux-mêmes (voir table 4). On peut ainsi aisément identifier le fichier choisi grâce à la composition des noms de fichiers qui est conforme aux recommandations CEE-ESPRIT «SAM» N°2589. Le nom du fichier de description orthographique est identique à celui du fichier signal auquel il est associé, à la différence près que dans l'extension du nom du fichier signal, le caractère "S" est remplacé par le caractère "O").

D'autre part, à tout fichier de signal est associé un fichier de configuration (informations sur la fréquence d'échantillonage, niveaux de déclenchement, tête de signal, etc.). Le nom de ce fichier est identique à celui du fichier signal et porte le suffixe ".CFG".

Accès aux données

Il est possible d'accéder aux données de la base de deux manières différentes: soit à l'aide du logiciel GERSONS dont nous avons

parlé plus haut, soit directement sous DOS.

L'accès à l'aide de GERSONS permet de choisir spécifiquement la séquence phonétique que l'on souhaite étudier.

Sous DOS, on accède directement aux fichiers de la base de façon descendante. Les fichiers sont stockés dans des sous-répertoires et sont classés respectivement selon la langue du locuteur (français/anglais), la pression (atmosphérique/hyperbare), la situation (caisson/piscine), et la profondeur.

4-CONCLUSIONS ET PERSPECTIVES.

L'existence de PSH/DISPE permet de faire interagir: recherche fondamentale, recherche et développement et standardisation. Ces trois pôles d'intérêt peuvent respectivement s'exprimer de la façon suivante:

- Modéliser la parole hyperbare: la base de données permet d'organiser la recherche fondamentale de façon plus systématique et cohérente. Elle rend possible, en effet, l'accès à du matériel linguistique structuré dont les conditions de production sont connues. Elle permet aussi de vérifier les prédictions d'un modèle.

- Développer de nouveaux produits: La modélisation de la parole hyperbare rend possible le développement de systèmes permettant de corriger spécifiquement les déformations mises en évidence au cours des analyses acoustiques. La recherche et le développement seront d'autant plus efficaces que les données de la base PSH/DISPE prennent en compte un grand nombre de problèmes de communication en milieu hyperbare ainsi qu'un éventail assez large de situations de communication.

- Evaluer l'intelligibilité: la conception des corpus de la base de données permet de réaliser des tests d'intelligibilité spécifiques à certaines situations de communication pour différentes profondeurs avec des mélanges respiratoires variés. La mise en oeuvre de ces tests permet d'identifier les traits phonétiques les plus dégradés. Ils permettent de la même façon de constater quels traits sont les mieux restitués après le passage par un «décodeur». Le diagnostic réalisé doit permettre d'orienter les efforts de développement. Enfin, l'évaluation de l'intelligibilité doit permettre de proposer des normes de qualification (AFNOR, 1991) pour les matériels de communication.

La base de données PSH/DISPE doit permettre de constituer à terme une base de connaissances.

La réalisation de cette base de données de parole hyperbare est un bon exemple de la complémentarité qui peut exister entre recherche fondamentale et recherche appliquée.

REFERENCES

AFNOR (1991). "Communications vocales en milieu subaquatique et/ou hyperbare: principes fondamentaux d'évaluation des équipements," Norme S 31-116, Marchal ed., September, ISSN 0335-3931.

Caeroux, J. C., and Dolmazon, J. M. (1990). "PTS Software V 4.21. User's Manual," *Periodic report ESPRIT PROJECT* 2589 (SAM), June, ICP, Grenoble.

Carré R. et al. (1987). "La base de données des sons du français (BDSONS). Perspectives et développement," in *Actes des 16èmes Journées d'Etude sur la Parole*, edité par la société française d'acoustique, Hammamet, 335-337.

Crestel, J. et Guitton, M. (1987). "Un système pour l'amélioration des communications en plongée profonde," in *Actes du 11ème Colloque GRETSI*, Nice.

Duncan, G. and Jack, M. (1987). "The helium speech effect," *Aspect of Speech Technology*, M. Jack & J. Laver eds., Edinburgh University Press, 216-283.

Espesser, R. et Balfourier, O. (1989). "V.E.S. mode d'emploi," Laboratoire Parole et Langage, Rapport interne, I.P.A., Aix-en-Provence.

Fant, G., and Lindqvist, J. (1968). "Studies related to diver's speech. Air pressure and gas mixture effects on diver's speech," *STL-QSPR* 1, 7-17.

Foulard, P. et Sérignat, J. F. (1991) "GERSONS: Système de gestion de la Base de Données des Sons du français. Manuel utilisateur," *GRECO-PRC* "Communication Homme-Machine", Rapport interne, January, I.C.P. Grenoble.

Fourcin, A. J., Harland, G., Barry, W., and Hazan, V. (1989).*Speech Input and Output Assessment*, Ellis Horwood ed., New-York.

Griffiths, J. D. (1967). "Rhyming minimal contrasts: a simplified diagnostic articulation test," *J. Acoust. Soc. Am.* 42, 236-241.

Hollien, H. and Hicks, J. W., (1982). "Helium/Pressure effects on speech: updated initiatives for research," *Proceedings of the IEEE Acoustic Communication Conference*, Washington, 1-26.

House, A. S., Williams, C. E., Hecker, M. H. L. and Kryter, K. D. (1965). "Articulation-Testing Methods: consonantal differancia-

tion with a closed-response set," *J. Acoust. Soc. Am.* 37, 158-166.
Zeiliger, J., and Sérignat, J.F. (1990). "EUROPEC V 4.0: a speech database recording software. User's Guide," *Periodic report ESPRIT PROJECT* 2589 (SAM), September, I.C.P. Grenoble.

Remerciements:

La base de données de parole hyperbare a été réalisée grâce au soutien financier de la Mission interministérielle de la mer, du C.E.P.&M., de la région P.A.C.A. et de la D.R.E.T. Nous adressons nos remerciements à tous les plongeurs qui ont participé aux enregistrements et aux équipes de techniciens et d'ingénieurs qui nous ont offert leur concours au GISMER (Toulon) et à l'INPP. Enfin, un grand merci à l'équipe de l'I.C.P. Grenoble qui nous a entraîné à l'utilisation du logiciel EUROPEC.

Distribution de PSH/DISPE: le CDROM contenant la base de données est distribué par CEDROM Technologies, 30 Avenue de l'Observatoire, 75014 Paris.

D
-
SCD: DC
SNM: DUPUIS
SBN: Charles
SEX: M
DOB: 1959
HET: 1,90
WET: 89
NLN: Français
ACC: Standard
ETH:
EDL: Secondaire
SMK: non fumeur
PTH:

Table 1: fichier de caractéristiques des locuteurs.

SCD: start of conditions
RCC: 180PiscMN
VER: V3.0
VOL:
DIR:
SNB: 2
SBF: 01
SSB: 16
NCH: 1
LGG: 0
PCN: 1
SAM: 40000
MIN: micro_name char(10)
FLT: filter char(10)
PRO: ext_proc char(10)
ENV: Piscine
DPT: 180m
GAZ: Héliox
PPO: 570 mb
RO2: 3
RHE: 92,7
TGZ:
TAQ: 5
MSK: X-LITE
ECD: end of conditions

Table 2: fichier de conditions d'enregistrement.

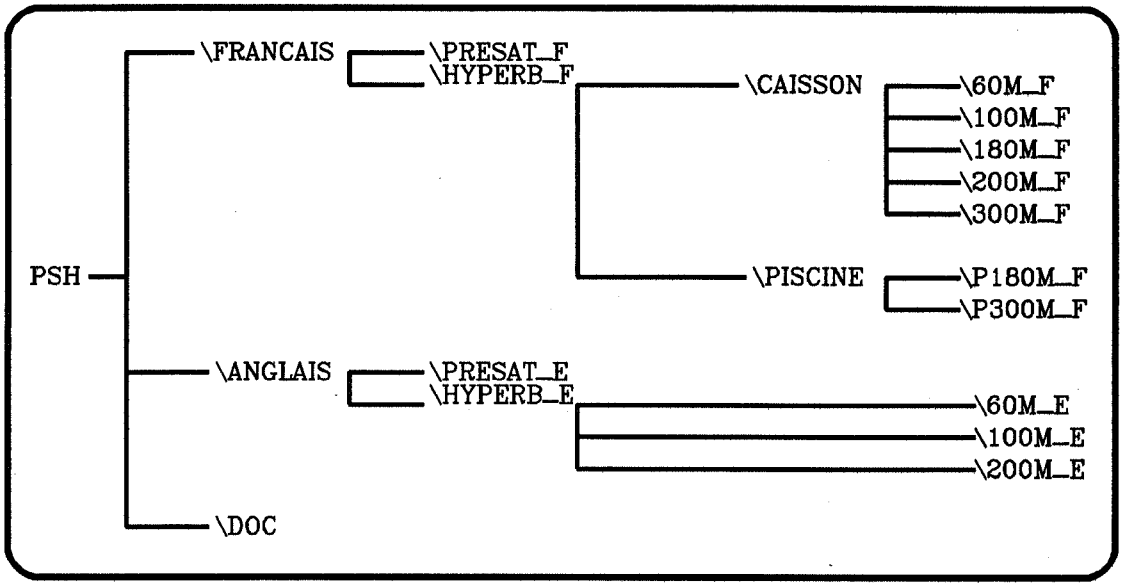
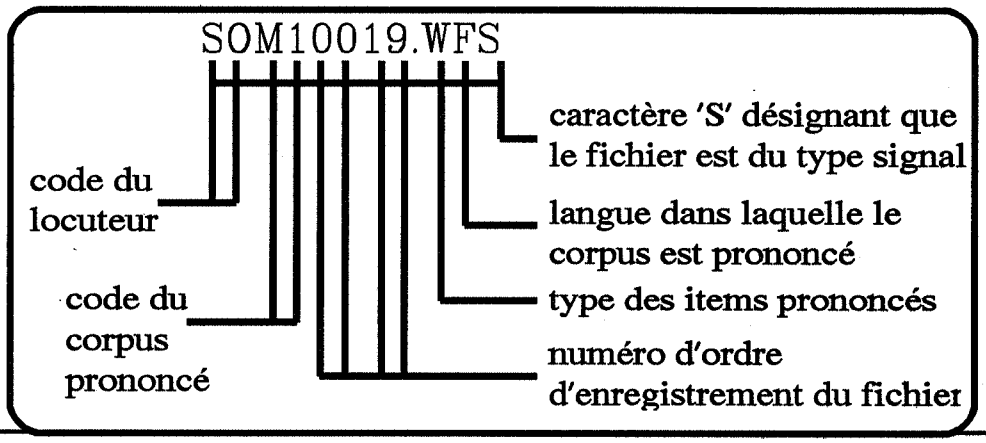


Table 3: Structure des répertoires et sous-répertoires dans le CDROM PSH/DISPE.



Conformément aux recommandations CEE-ESPRIT «SAM» N°2589:
1-Le premier caractère de l'extension du nom du fichier définit le type des items prononcés.
Soit : S : pour des phrases (Sentences),
P : pour un paragraphe (Passage),
W : pour des mots isolés (Isolated Word).
2-Le deuxième caractère de l'extension du nom du fichier indique la langue dans laquelle les items sont prononcés. Soit : E : pour l'anglais,
F : pour le français,
3-Le troisième caractère de l'extension du nom du fichier indique le type du fichier, soit ici «S» pour fichier de signal. Les différents types de fichier concernant PSH/DISPE et considérés par la norme SAM, sont désignés par les caractères suivants:
S : pour fichier d'échantillons de signal,
O : pour fichier de description orthographique.

Table 4: structure et signification des noms de fichier.

19es J.E.P. - Bruxelles, 19 au 22 mai 1992.

DEUX APPROCHES DE L'ETIQUETAGE EN EVENEMENTS PHONETIQUES

R. ANDRE-OBRECHT, G. PERENNOU, N. VIGOUROUX

IRIT - UA CNRS 1399 - UNIVERSITE PAUL SABATIER
118, route de Narbonne - 31062 Toulouse Cedex

Résumé

Cette communication a pour objet la comparaison de deux méthodes de segmentation du signal de parole en unités infra-phonétiques; elles s'opposent de par les paramètres utilisés (temporel et fréquentiel) et de par le type d'approche. L'une relève d'un système à base de connaissances et d'un étiquetage centiseconde ; l'autre est une technique empruntée au traitement du signal .

Les expériences, réalisées sur le corpus francais d'EUROM-0, ont pour but de montrer quels types d'unités peuvent être trouvés avec un minimum d'information (2 à 3 paramètres) et de quelle nature est cette information pertinente. La comparaison des unités obtenues avec un étiquetage manuel, nous a permis de juger de la cohérence entre unités acoustiques et unités phonétiques, et de fournir des résultats quantitatifs plus précis.

1 Introduction

Le décodage acoustico-phonétique, qu'il intervienne dans un système de reconnaissance automatique de parole ou dans un module d'étiquetage automatique, se décompose traditionnellement en trois étapes: la paramétrisation du signal, la segmentation du signal vocal, la définition et l'identification des unités "acoustico - phonétiques".

Les méthodes employées à chacun des niveaux peuvent être très variées, mais les systèmes se distinguent essentiellement par leur réponse à la question: "doit-on segmenter le continuum qu'est le signal de parole?"

Cette question brûlante a fait l'objet de nombreuses discussions lors de l'étiquetage manuel de la base de données BDSONS [7]: chacun aimerait poser sur le signal acoustique une marque signalant la frontière entre deux phonèmes; malheureusement un phonème est une unité linguistique. De ce fait, seuls les sons peuvent être éventuellement repérés et localisés ; de plus, du fait de la forte variabilité, des phénomènes d'assimilation et de coarticulation, rares sont les instants précis synonymes du passage d'un son à l'autre ; la seule segmentation que l'on puisse "espérer", sans faire appel à un système complexe, est une segmentation de nature acoustique .

De nombreuses études ont eu pour objet la segmentation du signal de parole, mais l'emploi d'un tel pré-traitement se justifie essentiellement par l'utilisation qui en est faite a posteriori, alignement phonétique, reconnaissance, synthèse, codage...

De manière schématique, trois types d'approche peuvent être envisagés pour segmenter le signal de parole:

- une partition d'un espace de paramètres permet un étiquetage centiseconde; chaque changement d'étiquettes motive une nouvelle frontière,
- après une analyse centiseconde fournissant pour chaque trame un vecteur d'indices de nature souvent acoustique, une fonction de discontinuité est évaluée à partir de plusieurs vecteurs ; le franchissement par cette fonction d'un seuil ou la localisation d'un maximum fournit une frontière,
- une analyse statistique permet par l'intermédiaire d'une modélisation du signal de confronter deux hypothèses ; la décision sera interprétée comme une présence ou une absence de rupture .

Nous nous proposons d'examiner deux méthodes de segmentation : la méthode développée par l'IRIT [1] dans un but d'alignement phonétique, s'apparente

aux deux premières approches, tandis que la méthode de divergence forward - backward conçue à l'IRISA [3], dans un but de reconnaissance, relève de la troisième .

Les expériences, réalisées sur le corpus francais d'EUROM-0, ont pour but de montrer quels types d'unités peuvent être trouvés avec un minimum d'information (2 à 3 paramètres) et de quelle nature est cette information pertinente. La comparaison des unités obtenues avec un étiquetage manuel, nous a permis de juger de la cohérence entre unités acoustiques et unités phonétiques, et de fournir des résultats quantitatifs plus précis.

2 Méthode SED

La méthode SED ("Segmentation, Etiquetage et Discontinuités") développée à l'IRIT, est basée sur une analyse temporelle du signal de parole et un ensemble de règles qui permettent d'étiqueter chaque trame de signal en événements phonétiques, puis d'en effectuer un regroupement afin d'obtenir la segmentation recherchée .

2.1 Paramètres acoustiques

Pour rendre compte des contrastes acoustiques qui peuvent motiver des frontières phonétiques, sont choisis l'amplitude et le taux de passage à zéro, paramètres fréquemment utilisés par les phonéticiens pour l'étiquetage manuel . Ces paramètres sont calculés et prétraités avec précision en fonction d'une double exigence :

- d'une part, il faut limiter la sursegmentation qu'entraîneraient des discontinuités non pertinentes (par exemple à l'intèrieur d'une fricative ou d'une aspiration) ; un lissage est donc nécessaire .
- d'autre part, il faut, autant que possible, préserver, voire amplifier les discontinuités essentielles séparant des événements phonétiques .

2.1.1 Calcul de l'amplitude

Sur chaque trame t de 4ms sont calculés le maximum $X(t)$ et le minimum $x(t)$ qui sont ensuite lissés sur une fenêtre glissante de 2K+1 trames par:

$$\begin{cases} MX(t) = max_{t-K \leq k \leq t+K} X(k) \\ mx(t) = min_{t-K \leq k \leq t+K} x(k) \end{cases}$$

On en déduit une première amplitude lissée
$y(t) = \frac{1}{2}(MX(t) - mx(t))$
qui a l'inconvénient de présenter une distorsion systématique de 4K ms (figure 1) se manifestant par une anticipation sur les faces montantes et par un retard sur les faces descendantes du profil d'amplitude réelle.

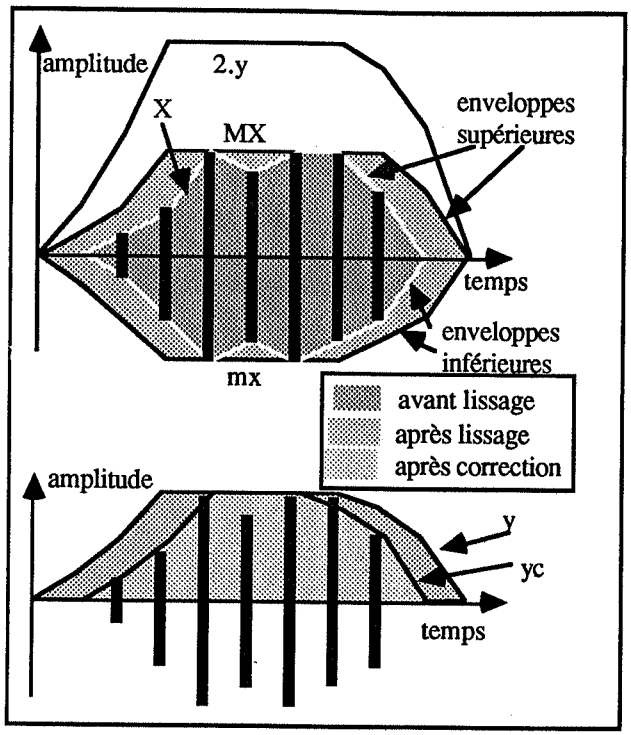


Fig.1 - Principe de calcul de l'amplitude.

L'amplitude corrigée s'obtient par :
$yc_t = min_{t-K \leq k \leq t+K} y(k)$
A ce stade des discontinuités isolées de l'ordre de la fenêtre de lissage sont absorbées (pour les résultats présentés, K=1) .

2.1.2 Lissage non linéaire de l'amplitude

Pour renforcer le lissage et en même temps renforcer les discontinuités essentielles nous procédons à un lissage suivi d'une correction visant à compléter l'élimination des instabilités isolées et à renforcer les fronts de montée ou de descente d'événements pertinents .

Le principe adopté est le suivant :

- pour chaque trame t une courbe lissée $b_t(j)$ est calculée sur la fenêtre
  $(t - J \leq j \leq t + J)$; (dans notre cas, J=3) ;
- on évalue pour chaque t,le meilleur modèle de lissage sur cette même fenêtre, appelé $b_s(.)$ (le critère permettant de juger du meilleur modèle peut être la meilleure approximation, la plus faible variance, ...) (figure 2) ;
- l'amplitude lissée est alors donnée par :
  $b_c(t) = b_s(t)$.

L'amplitude est ensuite convertie en unités logarithmiques puis normalisée entre [0,1] ; cette amplitude sera maintenant désignée A .

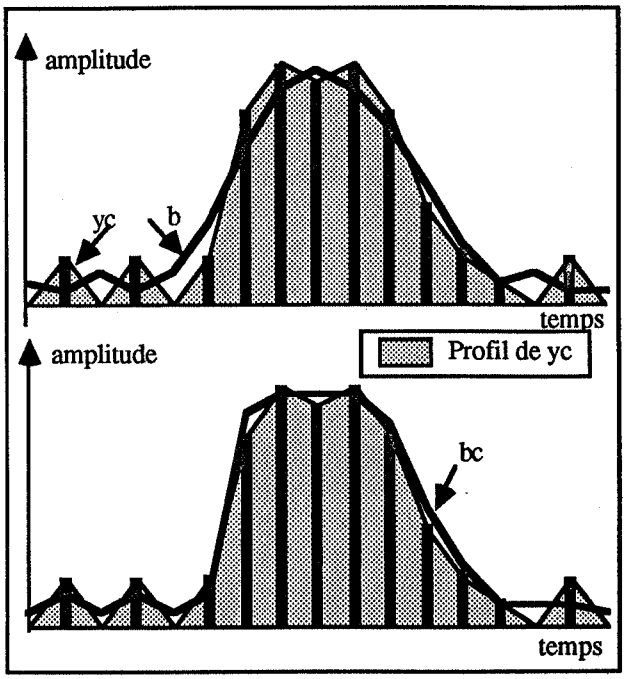


Fig.2 - Lissage et correction de l'amplitude. Le modèle local est ici la moyenne avec K=1. Le "meilleur modèle" de lissage est celui qui approxime le mieux yc.

2.1.3 Amplitude contextuelle et amplitude relative locale

Afin que l'amplitude discrimine au mieux les différentes classes de sons, tels que voyelles, sonantes, occlusives voisées et sourdes, il est nécessaire de la relativiser par rapport au niveau sonore de la voix. Nous introduisons l'amplitude contextuelle

$A_c(t) = max(S, A(n), t - N \leq n \leq t + N)$.

Dans notre cas, $N = 16$, de sorte que $A_c(t)$ est l'amplitude sur une fenêtre de 128ms centrée sur t; la valeur de S permet de lui conserver une valeur au dessus d'un seuil ; sans cette précaution, l'amplitude contextuelle tomberait à zéro dans les intervalles de pause, ce qui enlèverait toute signification à l'amplitude relative. Celle-ci est évaluée par :

$A_r(t) = \lambda A(t) + (1 - \lambda) \frac{A(t)}{A_c(t)}$

où λ est un paramètre compris entre 0 et 1 permettant une pondération.

2.1.4 Taux de passage à zéro

Calculé d'abord sur chaque trame de 4ms ce paramètre subit ensuite le même traitement que l'amplitude (ce paramètre donné en unité logarithmique et normalisé est désigné par Z) .

2.1.5 Indice d'instabilité

Il est basé sur l'évolution locale des paramètres A_r et Z et de leurs dérivées premières et secondes. On en déduit un indicateur qui est positionné à 1 si l'indice atteint un maximum au dessus d'un seuil, à 0 sinon .

2.2 Etiquetage en macro-classes phonétiques

Lorsque l'indice d'instabilité vaut 1, la trame est étiquetée A ou B selon que la pente de l'amplitude est croissante ou décroissante . Dans les autres cas, les étiquettes sont attribuées en fonction de Z et A . Le tableau 1 donne l'ensemble des 13 étiquettes utilisées .

Code	Signification
K	voyelle
N	voyelle aigu
M	voyelle grave
L	vocalique aigu
U	vocalique grave
O	occlusif voisé
Q	occlusif sourd
X	fricatif de type R
F	fricatif faible
Z	fricatif voisé
S	fricatif sourd
A	attaque discontinue
B	coda discontinue

Tableau 1 - Liste des événements phonétiques

Apres avoir regroupé les étiquettes de même nature, un ensemble de règles acoustiques fournit le dernier lissage et la segmentation recherchée en macro-classes . En particulier, les petits segments au contact de A et B peuvent être effacés par des règles de type :

$ATe \longrightarrow Ae$

où T est un petit segment et e une étiquette quelconque . Les segments A et B sont ensuite effacés par des règles de type :

$Ae \longrightarrow e$ *ou* $eB \longrightarrow e$.

Le résultat de ces traitements est de déplacer une partie des frontières vers les trames étiquetées A et B . L'ensemble de ce module d'étiquetage est écrit sous forme de fonctions de base et de réseaux dans l'environnement TEX [2].

3 Méthode de divergence Forward-Backward

La méthode de Divergence Forward-Backward, développée à l'IRISA, relève d'une étude statistique du signal. Elle a pour caractéristiques, un calcul séquentiel des paramètres à chaque échantillon, l'application d'un test statistique (modèle-test), et une détection séquentielle des ruptures.

Le signal (y_n) est supposé être décrit par une suite de zones quasi-stationnaires; chacune est caractérisée par un modèle auto-régressif gaussien :

$$\begin{cases} y_n = \Sigma a_i y_{n-i} + e_n \\ \text{var } (e_n) = \sigma^2 \\ \text{où } (e_n) \text{ est un bruit blanc gaussien.} \end{cases}$$

Le problème se réduit à détecter desdont changements dans les paramètres du modèle qui sont :

$$(A^T, \sigma) = (a_1, \cdots, a_p, \sigma)$$

Deux modèles $M_0(A, \sigma)$ et $M_1(A, \sigma)$ sont identifiés sur deux fenêtres d'analyse distinctes à chaque instant, une fenêtre croissante et une fenêtre glissante. Les deux fenêtres se superposent conformément à la figure 3.

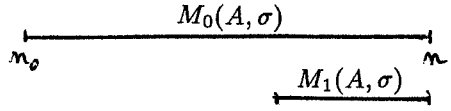


figure 3 : Localisation des modèles

Intuitivement, si aucun changement dans les paramètres n'intervient entre l'instant n_0 (correspondant à la dernière rupture détectée) et l'instant courant n , les modèles sont identiques et toute distance entre eux doit être nulle.

La distance statistique utilisée est obtenue à partir de l'entropie mutuelle des deux lois conditionnelles correspondantes, calculée à chaque instant k,

$$w_k = \frac{1}{2}\left\{2\frac{e_k^0 e_k^1}{\sigma_1^2} - \left[1 + \frac{\sigma_0^2}{\sigma_1^2}\right]\frac{e_k^0}{\sigma_0^2} + \left[1 - \frac{\sigma_0^2}{\sigma_1^2}\right]\right\}$$

avec

$$e_k^i = y_k - \sum_{j=1}^{p} a_j^i y_{k-j}, \quad i = 0, 1$$

erreur de prédiction à l'instant k pour chacun des modèles $M_i\,(A_i, \sigma_i)$; le test statistique est la somme cumulée sur le temps de ces valeurs . Le problème de détection de ruptures se résume à détecter un changement de moyenne . Il est résolu en utilisant la règle de Page-Hinkley qui consiste à ajouter un biais constant δ ; il suffit alors de détecter les maxima locaux moyennant un certain seuil λ.

Plus précisément, si on note :

$$\hat{W}_n = \sum_{k=1}^{n} (w_k + \delta)$$

une rupture est détectée à l'instant n_D définie par

$$n_D = \inf\left\{n \geq 0 \left| \max_{1 \leq k \leq n} \hat{W}_k - \hat{W}_n > \lambda \right.\right\} ;$$

la rupture a lieu à l'instant r, argument de l'expression: $\max_{1 \leq k \leq n_D} \hat{W}_k$.

Il est apparu pertinent pour fixer le biais et le seuil de distinguer deux zones, les zones voisées et les zones non voisées ; deux familles de (biais-seuil) sont donc introduites . De facon pratique, un test rudimentaire calculé à partir de l'énergie et du premier coefficient de réflexion durant la phase d'initialisation de chaque segment, permet de décider avec quels biais et seuil la statistique sera activée ; ce couple est alors inchangé tant qu'une nouvelle rupture n'est pas détectée.

Certaines frontières sont cependant systématiquement omises; par exemple, la frontière entre les sons [e m] est omise alors que celle entre [m e] ne l'est pas. Pour remédier à cet inconvénient la statistique est activée dans le sens rétrograde du signal pour tout segment estimé trop long. Cette procédure complète est la méthode de divergence forward-backward. (Pour plus de détails sur cette méthode, et sa mise en oeuvre, se reporter à [3]).

4 Protocole expérimental

4.1 Adaptation etApprentissage

Avant de procéder au traitement il est nécessaire de donner la composante continue du signal et un facteur de gain afin de s'adapter aux conditions d'enregistrement .

Dans la méthode SED, il est possible d'affiner les résultats en précisant des valeurs données a priori, à savoir la pré-emphase, le rapport signal/bruit . La mise au point des règles acoustiques a été faite à partir des connaissances d'un expert humain, et des resultats d'expérience sur une phrase issue du corpus francais de la base européenne EUROM-0 (locuteur DP).

Les paramètres nécessaires à la mise en oeuvre de la méthode de divergence sont la longueur de la fenêtre d'initialisation, l'ordre des modèles et les valeurs des deux familles biais-seuil. Apprises en expérimentant la méthode sur trois phrases issues d'une liste de phrases phonétiquement équilibrées, monolocuteur (signal fourni par le CNET Lannion), les variables sont fixées à 20ms pour la longueur de la fenêtre glissante, à 2 pour l'ordre du modèle, et le couple biais-seuil est donné par :

$$\begin{cases} (\delta_v, \lambda_v) = (0.2, 40) & \text{pour les zones voisées} \\ (\delta_b, \lambda_b) = (0.8, 80) & \text{sinon.} \end{cases}$$

D'autres valeurs seront testées afin de juger de la pertinence de l'information extraite.

4.2 Evaluation

Afin d'évaluer quantitativement la segmentation issue des deux méthodes proposées, nous avons comparé la localisation des frontières obtenues, à l'étique-

tage manuel effectué par des experts phonéticiens. Nous avons introduit deux critères, à savoir le taux de sur-segmentation et un critère dit de qualité.

- Le taux de sur-segmentation, noté TSS, est le rapport entre le nombre d'unités trouvées automatiquement et le nombre d'unités phonétiques identifiées par l'expert.
- Le coefficient de qualité de segmentation, noté CQ(δ), est le quotient par le nombre de frontières manuelles du nombre de ces frontières approximées par des frontières automatiques à δ ms près, en respectant la contrainte que deux frontières manuelles ne peuvent être approximées par une même frontière automatique.

5 Synthèse des résultats

L'évaluation des deux méthodes est faite sur la base de données EUROM-0. Dans cette communication, seuls des sous ensembles des corpus francais et anglais sont testés par chacune des deux méthodes .

5.1 Quelques chiffres

Le tableau 2 offre un résumé des expériences réaliseés à ce jour . Il apparait que le critère de qualité le plus élevé est obtenu par la méthode SED ; pour δ = 20ms, le critère de qualité est de 0,95 pour la méthode SED contre 0,92 pour la méthode de divergence . Par contre, le taux de sur-segmentation est plus élevé pour la méthode SED : il est de l'ordre de 2,5 contre 1,87. Notons qu'à 30ms, le critère de qualité est de 0,97 (resp. 0,96) pour la méthode SED (resp. pour la méthode de divergence).

	Méthode Divergence			Méthode SED		
Locuteurs	TSS	CQ(20)	CQ(30)	TSS	CQ(20)	CQ(30)
DP	1.85	0.937	0.960	2.43	0.951	0.971
SA	1.79	0.895	0.953	2.55	0.948	0.977

Tableau 2 - Résultats d'évaluation des deux méthodes

Une version préliminaire de la méthode SED a été évaluée sur les corpus anglais, francais et italien [2], sans recours à un nouvel apprentissage ; les résultats sont confirmés, ce qui prouve la robustesse de la méthode face aux différentes langues.

Plusieurs expérimentations ont été conduites sur ces données avec la méthode de divergence en faisant varier la longueur de la phase d'initialisation et l'ordre des modèles. Les résultats relatifs au critère de qualité sont identiques, que la longueur soit ramenée à 12ms ou que l'ordre des modèles passe de 2 à 16. Par contre, si la fenêtre est trop diminuée (tout en restant convenable par rapport à l'ordre des modèles identifiés) ou si l'ordre est élevé, le taux de sur-segmentation augmente sensiblement (+0,1). Il apparait, que, quelle que soit l'approche, l'information utile pour obtenir une segmentation acoustique correcte ($CQ(\delta) > 0,90$) se réduise à deux ou trois paramètres . L'ordre 2 utilisé pour l'identification des modèles (3 paramètres a_1, a_2, σ) laisse supposer que cette information est contenue dans le centre de gravité du spectre et les variations d'intensité, ce qui rejoint l'interprétation des deux paramètres temporels (l'amplitude et le taux de passage par zéro) .

5.2 Quelques commentaires

A l'issue d'un examen minutieux des frontières manuelles non approximées, il est difficile de dire qu'une méthode privilégie plus particulièrement un type d'erreurs . Par contre il est clair qu'un certain nombre de fautes proviennent de désaccords entre l'étiquetage manuel proposé et un étiquetage basé sur une analyse spectrale . Afin de nuancer nos propos, signalons que l'étiqueteur manuel avait pour but de localiser et délimiter tous les phonèmes et phénomènes extra-linguistiques, selon certaines normes et au vu du signal seul [5].

Les erreurs peuvent être classées en quatre catégories :

- des omissions nettes et précises où deux phonèmes sont clairement regroupés en une seule unité, alors que temporellement comme spectralement la frontière existe; citons pour exemple les sons [a l], [n y],...
- des délais supérieurs à 20ms :
  - pour les sons [s i], la structure formantique de la voyelle apparaît très tôt et se superpose au bruit de friction haute fréquence, d'où une avance à la détection ;
  - les frontières du son [R] en contexte vocalique sont difficiles à préciser, d'autant plus lorsque la pince formantique s'allonge
  - les frontières entre les semi- voyelles et voyelles peuvent être très floues ;
  - les frontières nasales-plosives posent problème, (figure 4) .
- dans la mesure où l'étiquetage manuel utilisé avait pour but de localiser tous les phonèmes, des omissions sont dues à des frontières indétectables ; citons le cas des sons [R l], [t R], on parle très souvent de "cluster" en reconnaissance;

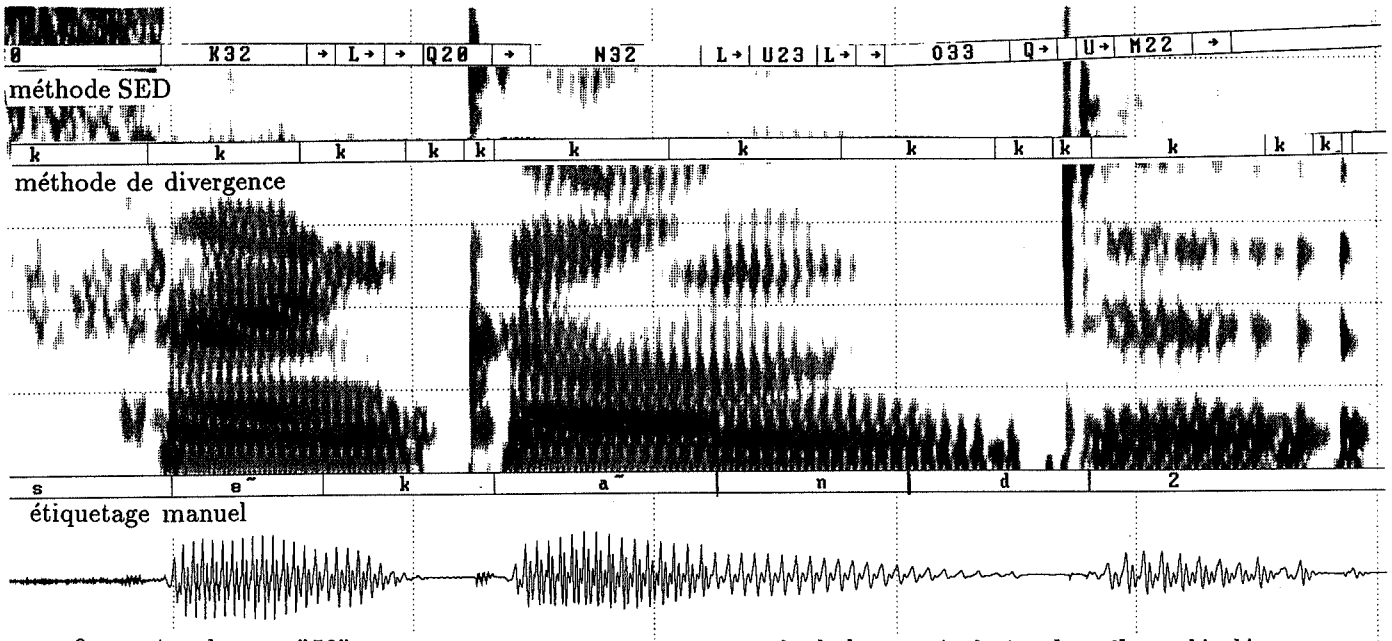


figure 4 : le mot "52" est prononcé, la réalisation spectrale de la coarticulation [ã t d] est décalée par rapport à l'étiquetage manuel.

- de manière plus rare, il y a complet désaccord; citons comme exemple le mot "dernier" qui a pour transcription normative [d e R J e] et dont la réalisation acoustique peut être [d e R n i] ; le désaccord porte sur la frontière [J e] .

Ces remarques ne remettent pas en cause l'étiquetage manuel mais elles mettent en évidence les difficultés d'évaluation des méthodes de segmentation.

6 Conclusion

L'évaluation des deux méthodes de segmentation, la méthode SED et la méthode de divergence, par rapport à un étiquetage manuel a permis de juger de la pertinence des frontières et d'interpréter les "erreurs".

Les deux méthodes mettent en évidence que peu d'informations (2 à 3 paramètres) sont nécessaires pour obtenir une segmentation utilisable en reconnaissance comme en alignement . Leurs résultats sont tout à fait comparables ; Les unités obtenues sont de nature infra-phonétique, (de l'ordre de 2 segments par phonème) . Peu d'erreurs graves sont observées mais la sur-segmentation est inévitable; elle facilite l'étape suivante, à savoir la paramétrisation des unités pour la phase d'alignement ou de reconnaissance . La paramétrisation d'unités de nature phonétique serait complexe : il est impensable de paramétrer globalement une plosive sans chercher où se réalise l'explosion! Cette étude est poursuivie afin de préciser le taux d'informations nécessaires à l'alignement et la reconnaissance [4,6]; la réponse ne devrait pas être alors la même...

Bibliographie

[1] H. KABRÉ, G. PÉRENNOU, N. VIGOUROUX, Automatic labelling of speech into events.*ICPhS,Aix en Provence*, august 1991.

[2] H. KABRÉ, Décodage acoustico-phonétique multilingue: système à bases de connaissances et étiquetage automatiqe de corpus de parole. *Thèse de doctorat de l'université Paul Sabatier de Toulouse*, septembre 1991.

[3] R. ANDRÉ-OBRECHT, A new statistical approach for the automatic segmentation of continuous speech signals. *IEEE Trans. on ASSP*, vol.36, no1, january 1988.

[4] R. ANDRÉ-OBRECHT, Reconnaissance automatique de parole à partir de segments acoustiques et de modèles de Markov cachés. *XVIII ièmes JEP Montréal*, juin 1990.

[5] D. AUTESSERRE ET C. MEUNIER, Segmentation criteria and labelling conventions. *Esprit Project 2589 (SAM), Multilingual speech input output*, decembre 1990.

[6] A. FARHAT, G. PÉRENNOU, N. VIGOUROUX, Segmentation en événements phonétiques et modèles markoviens pour l'étiquetage phonotypique. *XIX ièmes JEP Bruxelles*, mai 1992.

[7] L. MICLET , Présentation de la commission "Etiquetage Large" du GRECO Communication Parlée,*XVI ièmes JEP Hammamet*, octobre 1987.

19es J.E.P. - Bruxelles, 19 au 22 mai 1992.

LA COMPOSANTE PHONOGRAPHIQUE DE BDLEX

J.-M. PECATTE, M. de CALMES, D. COTTO, I. FERRANE, G. PERENNOU

IRIT - URA 1399 Université Paul Sabatier
118, route de Narbonne - 31062 Toulouse cedex

Résumé

Dans cet article, nous présentons la composante phonographique de la Base de Données et de connaissances LEXicales du français écrit et parlée (BDLEX). Cette composante est développée pour répondre aux besoins des applications mettant en jeu les relations oral/écrit de la langue comme la synthèse à partir de textes, la vérification et la correction lexicale de textes.

Ces travaux nous ont permis de recenser les phonogrammes fréquents décrivant la base de notre orthographe, aussi bien qu'un grand nombre de phonogrammes peu fréquents sur lesquels portent le plus souvent les difficultés de conversion graphème-phonème ou les erreurs orthographiques.

Après une description succincte de BDLEX, nous présentons une méthode de segmentation automatique en phonogrammes et les résultats obtenus sur le lexique de 23 000 entrées.

1. INTRODUCTION

Le projet BDLEX de Base de Données Lexicales du français écrit et parlé a été créé en 1983 dans le cadre du GRECO Communication Parlée (actuellement Pôle Parole du GDR-PRC Communication Homme-Machine) [Pérennou, 88]. Initialement destiné aux applications du traitement automatique de la parole, BDLEX s'est ensuite révélé utile dans d'autres domaines comme la vérification et la correction automatique de textes.

Pour traiter de manière efficiente les fautes phonographiques qui résultent d'une mauvaise transcription d'un ou plusieurs sons, les correcteurs doivent connaître les diverses associations lettres/sons et les erreurs observées dans le choix du code graphique [Pérennou, 86], [Véronis, 88].

Ces associations lettres/sons interviennent également dans la constitution d'une base de règles pour la transcription graphèmes-phonèmes en synthèse de la parole à partir de texte —pour un examen complet de ces questions on peut consulter [Sorin, 91].

Pour répondre efficacement aux besoins de ces applications, l'ajout d'une composante phonographique dans BDLEX s'est avéré nécessaire. Nous présentons, dans cet article, la méthodologie adoptée pour déterminer les associations lettres/sons ainsi que les résultats obtenus sur le lexique de BDLEX version 1 soit 23 000 entrées.

2. RELATION GRAPHIE-PHONIE-POSITION

2.1 LES GRAPHEMES ET PHONOGRAMMES

En pratique, aucune langue ne respecte, ce qui serait l'idéal dans le domaine, la correspondance terme à terme entre les graphèmes et les phonèmes. Par exemple, en français :
(1) - le graphème *ch* se prononce [ʃ] (chaud) ou parfois [k] (chorale)...
(2) - le son [k] correspond au graphème *c* (cas), *qu* (qui) ou *ch* (chorale) ...

Des associations lettres/sons, notées (G,S), telles que celles indiquées dans (1) et (2) sont appelées phonogrammes [Catach, 84].

La composante graphique G d'un phonogramme est appelée graphème. Certains graphèmes sont sensibles au contexte : si on laisse de côté les mots d'origine étrangère et quelques autres comme *amygdale*, g se prononce [ʒ] en contexte *__e*, *__i*, *__y* et [g] dans les autres cas, ce qui donne lieu à quatre phonogrammes :

(g,ʒ), (g,g), (gu,g) et (ge,ʒ) —les deux derniers apparaissant par exemple dans des mots comme *baguette* ou *bougeoir*.

En général, un seul son est associé à un graphème. Il existe cependant quelques exceptions : ainsi le h muet ne correspond à aucun son tandis que la lettre x peut se prononcer [ks] ou [gz].

Dans certaines applications comme la correction automatique de fautes d'orthographe, la notion de phonogramme telle que nous venons de la définir est imprécise puisque entre autres elle ne tient pas compte de la position dans le mot. Or la distribution des fautes phonographique est au contraire sensible à cette position : par exemple une faute de doublement de consonne s'observe rarement en début de mot ou en finale absolue (*basilic* ne se réécrit pas **bbasilicc*).

C'est pourquoi nous avons introduit le phonogramme positionnel (ou p-phonogramme) qui se présente comme un triplet (G,S,P) où P peut prendre les valeurs : D (début de mot), M (milieu), F (fin de mot). Dans les mot composés d et f désignent respectivement le début et la fin de mots autres que D et F.

2.2. DONNEES CLASSIQUES SUR LES PHONOGRAMMES DU FRANÇAIS

Des études récentes sur le système phonographique du français ont été réalisées pour des applications relevant du domaine de l'enseignement de l'orthographe et de la correction lexicale.

S'il s'agit de l'enseignement de l'orthographe, l'étude porte sur les phonogrammes les plus fréquents de la langue. N. Catach a proposé une description du système phonographique de base : le plurisystème du français [Catach, 84].

Il faut cependant observer que, pour des usagers ayant un bonne connaissance de la langue, les difficultés orthographiques portent plus vraisemblablement sur les phonogrammes peu fréquents pour lesquels on possède peu de données quantitatives.

Dans le cadre de la correction lexicale, J. Véronis a recensé les associations lettres/sons sur un lexique de 3274 mots parmi les plus fréquents du français : l'Echelle Dubois-Buyse d'Orthographe Usuelle Française. Il dénombre 93 graphèmes qui correspondent à 141 phonogrammes [Véronis, 88]. Cette étude a permis de présenter des fréquences comparables à celles que l'on trouverait dans l'analyse d'un corpus de texte. Cependant ces travaux sur un lexique de 3274 mots sont insuffisants.

C'est pourquoi nous avons recherché la disttribution des phonogrammes dans le lexique de BDLEX version 1 qui comporte 23 000 entrées. Cela nous a permis de recenser les phonogrammes fréquents décrivant la base de notre orthographe aussi bien qu'un grand nombre de phonogrammes peu fréquents.

3. LES PHONOGRAMMES et BDLEX

Après une description succincte de BDLEX, nous présentons la méthode de segmentation automatique en phonogrammes et les résultats obtenus sur le lexique de BDLEX version 1.

3.1. DESCRIPTION DU LEXIQUE DE BDLEX

Les informations associées à une entrée lexicale concernent :
- la graphie,
- la phonologie incluant une indication de syllabation,
- la morphologie flexionnelle (conjugaison des verbes, flexion des noms adjectifs),
- la morphologie dérivationnelle,
- la morphosyntaxe,
- la fréquence d'apparition dans les textes.

Seules les informations référencées par la composante phonologique sont décrites ici. Des informations complémentaires sont données dans [Pérennou,88]

Le champ PHON_SYLL fournit la représentation phonologique en syllabes et en pieds de l'entrée lexicale. Au niveau du lexique, les variantes des voyelles à double timbre sont seulement distinguées dans les cas du /ɔ/, du /o/, du /œ/ et du /ø/ en contexte de syllabe fermée accentuée.

Les segments consonantiques possèdent un statut fixé par un diacritique phonologique qui peut être l'un des suivants :
- fixité : c'est le statut par défaut ; ex : les consonnes de «bec» représenté par /bɛk/
- latence : on écrit X" si X ne se réalise qu'au contact d'une voyelle placée après et dont elle n'est séparée au plus que par une frontière faible entre mots (ex: le t final de «petit» —> /pəti t"/).
- mixité : X' si X est fixe devant une frontière forte et sinon latent (ex: «huit» —> /*ɥi t'/ ou * est mis pour h disjonctif, c'est-à-dire: interdisant la liaison).

A ces trois diacritiques —absence de marque, " et '— nous en ajoutons trois autres pour traduire des variantes interindividuelles dans les représentations lexicales :

X˙ : jamais prononcé par certains, toujours prononcé par d'autres (ex: «ananas» —> /anana s˙/)

X¯ : fixe ou mixte selon le locuteur (ex: «cinq» —> /sɛ̃ k¯/)

X⁺ : latent ou mixte selon le locuteur (ex: «suspect» —> /syspɛ̃ kt÷/ (ici c'est le groupe kt qui est latent ou mixte).

Ce sont les phonèmes terminaux qui sont sujets à ces variations. C'est pourquoi dans BDLEX un champ finale phonologique FPH a été prévu pour en faciliter la représentation. Il contient les finales ayant un statut autre que que celui de phonème fixe ; en plus des cas examinés précédemment, il contiendra aussi le schwa terminal.

3.2. SEGMENTATION AUTOMATIQUE EN PHONOGRAMMES

Pour effectuer une segmentation automatique des entrées de BDLEX en phonogrammes, nous avons adopté une méthode d'alignement entre la graphie accentuée et la représentation phonologique d'un mot. Elle est réalisée par le système VERITEXT qui utilise un ensemble de règles stochastiques décrivant les relations graphie-phonie de la langue. L'algorithme de segmentation utilisé dans VERITEXT est dérivé de celui du système VERIPHONE, système d'alignement phonétique [Pérennou, 89].

Les critères de segmentation utilisés respectent le principe qui est de constituer des regroupements entre lettres et sons se ramenant aussi souvent que possible à des phonogrammes attestés, c'est-à-dire présents dans de nombreux mots de la langue.

a m y g d a l e A m i ∧ d A l ə	**i c e b e r g** Aj s ə b ɛ r g
a x e A ks ə	**t o c s in** t o k s ɛ̃
b a h u t b A ∧ y ∧	**h o mm e** ∧ ɔ m ə
t i ck e t t i k e ∧	**aoû t** u t.
sc i en c e s j ã s ə	**e s c a l e** ɛ s k A l ə

Fig 1 - Exemples de segmentation en phonogrammes

3.3. RESULTATS SUR LES PHONOGRAMMES

La segmentation automatique des entrées lexicales de BDLEX nous a permis de recenser 350 phonogrammes [Pécatte, 92].

Ce nombre peut paraître important. En effet selon [Catach,84], la base de notre orthographe peut être décrite à l'aide d'une centaine de phonogrammes.

Mais dès que l'on considère un lexique suffisamment étendu, comme dans notre cas, il faut tenir compte des mots d'origine étrangère et des termes techniques qui contiennent un grand nombre de phonogrammes peu utilisés dans la plupart des mots fréquents.

Par exemple pour le son /k/, treize transcriptions ont été dénombrées (voir le tableau ci-après —à chaque graphème, entre parenthèses sa fréquence d'apparition dans BDLEX).

/k/	c	(5093)	*truculent*
	qu	(1329)	*quatre*
	k	(214)	*fakir*
	ch	(168)	*chorale*
	cc	(115)	*accord*
	ck	(62)	*ticket*
	cqu	(21)	*acquérir*
	q	(12)	*coq*
	kh	(12)	*cheikh*
	lk	(7)	*talkie-walkie*
	cch	(5)	*bacchanale*
	kk	(2)	*drakkar*
	ckh	(1)	*blockhaus*

De plus, contrairement aux travaux mentionnés, nous avons pris en compte les variations phonologiques en finale, ce qui nous a amené à créer 40 phonogrammes supplémentaires. Ainsi le statut de la consonne *d* en finale donne lieu aux quatre phonogrammes—(d,∧),(d,d),(d,d"),(d,d')— illustrés par les exemples suivants :

accord	(a,A)(cc,k)(o,ɔ)(r,r) **(d,∧)**
bled	(b,b)(l,l)(e,ɛ) **(d,d)**
allemand	(a,A)(ll,l)(e,ə)(m,m)(an,ã) **(d,d")**
baroud	(b,b)(a,A)(r,r)(ou,u) **(d,d')**

On notera que des phonogrammes tels que (d,d") contiennent en fait une information morphologique. Ils sont à rapprocher des morphogrammes introduits dans N. Catach.

Nos phonogrammes tiennent également compte des variations des voyelles à double timbre lorsque celles-ci sont distinguées dans la représentation phonologique de BDLEX (on y distingue par exemple : /o/, /ɔ/ et /O/, ce dernier laissant la possibilité de prononcer [o], [ɔ] ou un son intermédiaire). Ainsi le graphème *au* peut avoir trois représentations phonologiques en fonction du contexte :

centaure	(c,s)(en,ã)(t,t)**(au,ɔ)**(r,r)(,e,e)
assaut	(a,A)(ss,s)**(au,o)**(t,∧)
chaudron	(ch,ʃ)**(au,O)**(d,d)(r,r)(on,õ)

Pour caractériser l'efficacité d'un ensemble de phonogrammes, on introduit le pourcentange de couverture lexicale qui se définit comme la proportion des mots du lexique n'utilisant que les phonogrammes de cet ensemble.

La figure 2 précise la couverture lexicale des phonogrammes : en abscisse est donné le nombre de phonogrammes, en ordonnée le taux de couverture

lorsque les phonogrammes sont choisis comme étant les plus fréquents.

On observe que les 80 phonogrammes les plus fréquents couvrent 95 % des mots du lexiques. Ils suffisent pour rendre compte de la base de notre orthographe.

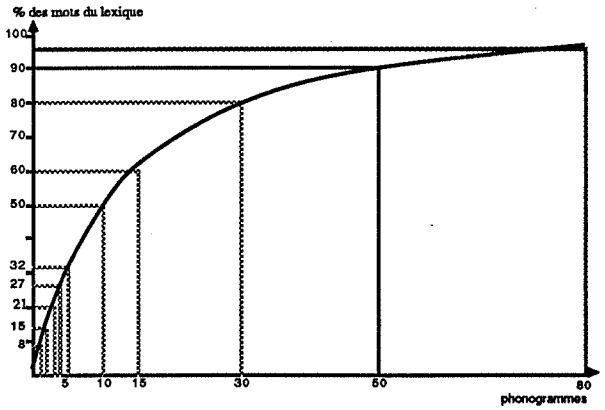


Fig 2. - La couverture lexicale des phonogrammes.

Les 270 phonogrammes restants ne sont utilisés que dans 5% des mots du lexique de BDLEX, soit 1150 mots, lesquels supportent la plupart des difficultés orthographiques.

Lorsque l'on examine les phonogrammes les moins fréquents, trois caractéristiques se dégagent :

- Certains contiennent des sons inhabituels pour le français dont la transcription dans le dictionnaire est une approximation grossière correspondant mal à la graphie, exemple : (aw,ɔ) dans *tomahawk.*

- D'autres contiennent des graphèmes inconnus du français de base, exemple : (ä,a) dans *hammän*, (ñ,ɲ) dans doña, (ooing,wɛ̃) dans *shampooing*.

- Enfin il peut s'agir d'associations inhabituelles de sons et de graphèmes courants, exemple : (u,o) dans *chewing-gum*, (j,r) dans *jota*, (o,u) dans *rahat-lokoum.*

Quelques mots courants contiennent aussi des phonogrammes rares, exemple : (aoû,u) dans *août*, (aon,ã) dans *faon.*

Souvent cependant un phonogramme non usuel caractérise l'origine linguistique du mot qui le contient. Ceci peut avoir des implications en synthèse multilingue où la détection de graphèmes particuliers peut aider à la détermination automatique de la langue d'un texte ou simplement d'un mot d'origine étrangère —pour une étude statistique de fréquences de suites de lettres dans différentes langues voir [Church,85].

3.4. RESULTATS SUR LES P-PHONOGRAMMES

Aux 350 phonogrammes correspondent 720 p-phonogrammes.

Ainsi le phonogramme (c,k) peut apparaître en début (D), milieu (M) ou fin (F) de mot ainsi qu'en début (d) ou fin (f) d'éléments de mots composés. Il donne lieu aux 5 p-phonogrammes suivants : (c,k,D), (c,k,M), (c,k,F), (c,k,d), (c,k,f). Le phonogramme (cch,k) se rencontre seulement en position interne, il lui correspond un seul p-phonogramme (cch,k,M).

Le tableau ci-après illustre les fréquences d'apparition dans le lexique de BDLEX des p-phonogrammes ayant le son [k].

S	G	total	M	D	d	F	f
[k]	c	5093	2476	2297	183	106	31
	qu	1329	11105	172	52	0	0
	k	214	101	84	4	22	3
	ch	168	102	51	10	5	0
	cc	115	115	0	0	0	0
	ck	62	23	0	0	28	11
	cqu	21	21	0	0	0	0
	q	12	7	0	1	3	1
	kh	12	5	6	0	1	0
	lk	7	7	0	0	0	0
	cch	5	5	0	0	0	0
	kk	2	2	0	0	0	0
	ckh	1	1	0	0	0	0

Dans le cadre de la correction lexicale, les statistiques sur les p-phonogrammes permettent de définir les erreurs observables dans le choix du code graphique. On peut ainsi déterminer automatiquement les réécritures fautives possibles d'un p-phonogramme en tenant compte le la fréquence d'apparition des p-phonogrammes ayant même valeur phonique et même position. Par exemple, les erreurs observables dans le choix du code graphique du p-phonogramme rare (kh,k,D) sont *c,qu,k* ou *ch* ; pour le p-phonogramme très fréquent (c,k,D), la seule réécriture fautive envisagée est *qu.*

Cependant, l'utilisation des p-phonogrammes n'est pas suffisante pour décrire toutes les erreurs observables : la difficulté orthographique dans tocsin ne porte pas sur un seul p-phonogramme mais sur le groupe de (c,k,M)(s,s,M) qui peut se réécrire fautivement x. C'est pourquoi, nous avons introduit, dans le système VORTEX, la notion de gpo (graphique-phonétique-orthographique) [Pérennou, 86] [Pecatte, 92].

3.5. RESULTATS SUR LES GRAPHEMES

Les résultats de la segmentation en phonogrammes sur le lexique BDLEX-1 ont fait apparaître 172 graphèmes.

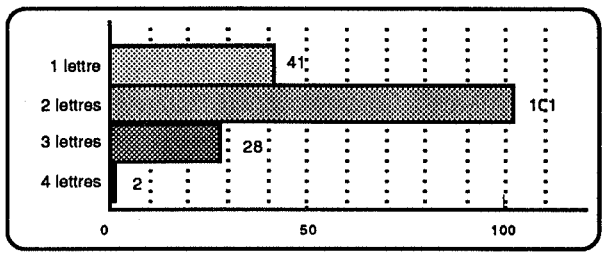


Fig.3. - La complexité des graphèmes

Ces graphèmes se répartissent en fonction de leur complexité graphique, comme indiqué sur la figure 3. Si on résonne en de fréquence lexicale, on peut observer que le système phonographique français est composé principalement d'éléments de longueur 2. La majorité des graphèmes (70 %) ont une seule représentation phonétique ; par exemple (an,[ɑ̃]), (m,[m]), (ph,[f]), parmi les plus fréquents. On peut aussi remarquer que la majorité des consonnes doubles *bb, dd, ff, kk, ll, mm, nn, pp, rr, ss, tt* fait partie de ce groupe, *gg* et *zz* faisant exceptions.

De même, presque tous les graphèmes ayant une fréquence d'apparition très faible ont aussi une seule représentation phonétique (c'est vrai évidemment pour ceux qui n'apparaissent qu'une seule fois).

Comme exception on peut noter que *ü* qui n'apparaît que dans deux mots se prononce [y] dans *führer* et [ɔ] dans *capharnaüm.* A l'opposé, le graphème *u* possède de nombreuses prononciations différentes :

u	[y]	*voluptueux*	[ɥ]	*écuelle*
	[ɔ]	*triclinium*	[œ]	*auburn*
	[Œ]	*trust*	[u]	*alléluia*
	[i]	*business*	[ju]	*trade-union*
	[w]	*cacahuète*		

4. LA COMPOSANTE PHONOGRAPHIQUE DE BDLEX

Chaque entrée lexicale de BDLEX contient sa représentation en phonogrammes.

La figure 4 montre quelques exemples de décomposition en phonogrammes. La représentation en p-phonogrammes d'une entrée lexicale peut être déduite implicitement de sa représentation en phonogrammes.

Le rôle d'une composante phonographique est de générer les représentations en phonogrammes des formes fléchies ou dérivées d'une entrée lexicale en tenant compte des variations morphophonologiques et phonologiques contextuelles. Cette composante est construite à partir des règles morphophonologiques et orthographiques.

La composante phonographique prend en compte les variations morphophonologiques et phonologiques contextuelles qui peuvent intervenir en finale de mots, dans le cas de flexion ou de liaison. On peut alors étendre cette composante sur la plan morphologique flexionnel et dérivationnel. Ainsi, il est possible de générer les représentations phonographiques des formes fléchies ou dérivées en adaptant, sur le plan phonographique, les règles flexionnelles et dérivationnelles existantes.

Par exemple, les règles flexionnelles qui permettent de générer les représentations graphique, phonologique et phonographique de la forme féminin singulier de *géant,* sont les suivantes :

Règle morpho-orthographique fém-e
géant + fém-e → géante
Règle morphophonologique fém-ə
ʒeɑ̃t"+ fém-ə → ʒeɑ̃te
moyennant la règle phonologique :
C" → C / __ [-cons]
Règle morphophonographique fém-(e,ə)
[(g,ʒ)(é,E)(an,ɑ̃)(t,t") + fém-(e,ə)
→ (g,ʒ)(é,E)(an,ɑ̃)(t,t)(e,ə)
moyennant la règle :
(X,C") → (X,C) / __ [-cons]

De même, les règles dérivationnelles qui permettent d'obtenir les représentations graphique, phonologique et phonographique de la forme préfixée par *in* de *réel,* sont les suivantes :

Règle morpho-orthographique $[in]_{préf}$
$[in]_{préf}$+[réel]$_{adj}$ → [irréel]$_{adj}$
moyennant la sous-règle : $[in]_{préf}$ → ir /_ r
Règle morphophonologique $[\tilde{\varepsilon}]_{préf}$
$[\tilde{\varepsilon}]_{préf}$+[reɛl]$_{adj}$ → [ireɛl]$_{adj}$
moyennant la sous-règle : $[\tilde{\varepsilon}]_{préf}$ → ir / _ r
Règle morphophonographique $[in,\tilde{\varepsilon}]_{préf}$
$[(in,\tilde{\varepsilon})]_{préf}$+[(r,r)(é,e)(e,ɛ)(l,l)]$_{adj}$
→ (i,i)(rr,r)(é,e)(e,ɛ)(l,l)
moyennant la sous-règle :
$[in,\tilde{\varepsilon}]_{préf}$ → (i,i)(r,r) / _ (r,r)

De telles règles jouent un rôle important dans la transcription graphème-phonème de qualité [Olive, 85].

5. CONCLUSION

L'étude que nous avons effectuée sur la structure phonographique de BDLEX a permis de recenser les phonogrammes (au total 350) que l'on peut rencontrer dans les textes réels.

L'ajout d'une composante phonographique permet de répondre plus efficacement aux besoins des applications du traitement automatique de la parole et de textes mettant en jeu les relations oral/écrit de la langue comme la synthèse de la parole, la vérification et la

correction de fautes d'orthographe. Les phonogrammes sont à la base des représentations lexicales utilisées dans notre correcteur VORTEX.

REFERENCES

[Catach, 84] N. Catach, *L'orthographe*, Collection Que sais-je ?, PUF, 1984.

[Church,85] K.N. Church, *Stress Assignment in Letter-to-Sound Rules for Speech Synthesis*, Proc. 23rd Meeting Ass. Comp. Ling., 246-53.

[Pécatte, 92] J.-M. Pecatte, *Tolérance aux fautes dans les interfaces homme-machine. traitement des chaînes phonétiques, des chaînes orthographiques et des structures syntaxiques*, Thèse de l'Université Paul Sabatier de Toulouse III, 29 janvier 1992.

[Pérennou, 86] G. Pérennou, P. Daubèze, F. Lahens, *La vérification et la correction automatique des textes : le système VORTEX*, Technique et Science Informatique, vol.4, n°5, pp.285-305, 1986.

[Pérennou, 88] G Pérennou, *Le projet BDLEX de base de données et de connaissances lexicales et phonologiques*, Actes des premières journées nationales du GRECO-PRC Communication Homme-Machine, EC2 Editeur, Paris, 24-25 novembre 1988.

[Olive,85] K.P. Olive, M.Y. Liberman, *Text-to-Speech : an Overview*, JASA, suppl. 1, 78-86.

[Pérennou, 89] G Pérennou, M. de Calmès, J.M. Pécatte, N. Vigouroux, *Phonetic-string alignment for an automatic labelling of speech corpora*, Proceedings of ESCA workshop on speech Input/Output Assessment and Speech Databases, The Netherlands, paper 5.4., September 20-23 1989.

[Sorin,91] C. Sorin, *Synthèse de la parole à partir du texte : «Etat des recherches et des applications»*, Actes de 2èmes journées nationales du GDR PRC Communication Homme-Machine, Toulouse, 29-30 janvier 1991.

[Véronis, 88] J. Véronis, *Le traitement de l'erreur dans le dialogue homme-machine en langage naturel*, Thèse de l'Université d'Aix-Marseille III, octobre 1988.

GRAPHIE	PHON_SYLL	FPH	CS	PHONOGRAMMES
aigre-doux	E/grə/du	s"	J	(ai,E) (g,g) (r,r) (e,ə) (-,∧) (d,d) (ou,u) (x,s")
amygdale	A/mi/dAl	ə	N	(a,A) (m,m) (y,i) (g, ∧) (d,d) (a,A) (l,l) (e,ə)
axe	Aks	ə	N	(a,A) (x,ks) (e,ə)
bahut	/bA/y		N	(b,b) (a,A) (h, ∧) (u,y) (t,∧)
chat-huant	/ʃA/y/ã		N	(ch,ʃ) (a,A) (t,∧) (-,∧) (h,∧) (u,y) (an,ã) (t,∧)
dix-huit	/di/zɥi	t'	J	(d,d) (i,i) (x,z) (-,∧) (h,∧) (ui,ɥi) (t,t')
iceberg	Ajsə/bɛrg		N	(i,Aj) (c,s) (e,ə) (b,b) (e,ɛ) (r,r) (g,g)
géant	/ʒE/ã	t"	N	(g,ʒ) (é,E) (an,ã) (t,t")
hautbois	/*O/bwA		N	(h,*) (au,O) (t,∧) (b,b) (oi,wA) (s,∧)
homme	ɔm	ə	N	(h,∧) (o,ɔ) (mm,m) (e,ə)
onze	/*õz	ə	J	(∧,*) (on,õ) (z,z) (e,ə)
tocsin	/tOk/sɛ̃		N	(t,t) (o,O) (c,k) (s,s) (in,ɛ̃)

Fig.4- BDLEX-1 Phonogrammes

19es J.E.P. - Bruxelles, 19 au 22 mai 1992.

ANALYSE LEXICALE DU CORPUS DE LA BASE DE DONNEES "BREF"

I. FERRANE, M. DE CALMES, D. COTTO, J.M. PECATTE, G. PERENNOU

IRIT - UA CNRS 1399 - UNIVERSITE PAUL SABATIER
118, route de Narbonne - 31062 Toulouse Cedex

Résumé

Dans cet article, nous présentons un ensemble d'observations statistiques faites sur le vocabulaire extrait d'un corpus de textes réels. Ceux-ci sont utilisés dans la base de données BREF, destinée au développement et à l'évaluation des *machines à dicter*.

Nous mettons en relief différents besoins lexicaux pour les traitements du français écrit et oral, comme la correction automatique de textes et le traitement automatique de la parole. Nous montrons la nécessité de développer des procédures de traitement des mots inattendus qui apparaissent très fréquemment dans les textes usuels. Ces procédures sont essentiellement basées sur l'étude de la structure morphologique des mots.

Pour réaliser cette étude, nous utilisons comme matériaux de référence, les données lexicales disponibles dans BDLEX. Cette base de données lexicales du français écrit et oral, a pour objectif de créer, d'organiser et de distribuer des matériaux lexicaux destinés au traitement automatique de la parole et des textes.

1. INTRODUCTION

Dans le domaine des *Industries de la Langue* les lexiques électroniques occupent une place importante. Dictionnaires et encyclopédies sont maintenant disponibles sous DOC ; pour le français, on peut citer entre autres le dictionnaire Zyzomis d'Hachette et le Robert électronique. Les systèmes de traitement de texte actuels disposent, dans leur environnement, de lexiques pouvant être consultés pour vérifier l'orthographe ou la conjugaison d'un mot, pour rechercher des synonymes, etc. Les correcteurs automatiques fonctionnent également en association avec des lexiques.

Actuellement, tous ces matériaux lexicaux sont encore loin de satisfaire les besoins du traitement automatique de la parole et des textes. Ils sont insuffisants lorsqu'on aborde des traitements linguistiques mettant en jeu une analyse sémantique et syntaxique approfondie. Ils sont également inadaptés pour des traitements de surface tels que ceux qui interviennent dans la correction orthographique, la synthèse de la parole à partir de texte, et la dictée vocale. C'est pourquoi, différentes équipes de recherche ont entrepris de développer leurs propres lexiques.

Dans cet article, nous décrivons les observations statistiques faites sur le vocabulaire extrait d'un corpus de textes réels constitué d'articles de journaux. Ceux-ci sont utilisés dans la base de données BREF, destinée au développement et à l'évaluation des *machines à dicter*.

Cette étude met en relief différents besoins lexicaux et montre la nécessité de développer tout un ensemble de procédures pour traiter les inattendus qui, quelle que soit l'étendue des matériaux lexicaux utilisés, surviennent très fréquemment dans les textes usuels.

Nous avons pris comme référence les matériaux disponibles dans la base de données lexicales du français écrit et oral, BDLEX, dont l'objectif est de créer, d'organiser et de distribuer des matériaux lexicaux destinés au traitement automatique de la parole et des textes. Pour plus de détails sur l'état actuel de BDLEX, on peut se reporter à [Pérennou, 92].

Les projets BREF et BDLEX sont développés dans le cadre du GDR-PRC Communication Homme-Machine —groupe de recherches coordonnées du Ministère de la Recherche et de la Technologie, et du Centre National de la Recherche Scientifique.

2. COUVERTURE DE TEXTES REELS

L'accès au lexique joue un rôle crucial pour des applications comme la correction automatique, et le traitement automatique de la parole. Si un mot est inconnu du lexique, le système est mis en échec sans qu'il le sache toujours. En effet, l'accès étant tolérant aux fautes ou aux imprécisions de reconnaissance, il se trouvera toujours un mot, plus ou moins proche, pour remplacer celui qui est observé.

Le *taux de couverture lexicale*, ou proportion des mots d'un texte connus du lexique, est donc un des critères importants pour l'évaluation du niveau de performance des systèmes de ce type. Cependant, il faut savoir que la nature des textes sélectionnés en vue d'être étudiés, peut conditionner qualitativement et quantitativement les résultats obtenus. En effet, ceux-ci seront différents selon le niveau du vocabulaire employé dans les textes : vocabulaire familier ou littéraire, général ou technique, ... [Catach, 84].

On distingue deux méthodes d'obtention du taux de couverture lexicale :

- la première prend en compte les mots par rapport aux textes ; ce qui signifie qu'à chaque forme rencontrée, peut être affecté un poids établi en fonction sa fréquence d'apparition dans le corpus de textes étudiés.
- la seconde, au contraire, est basée sur l'étude d'un lexique de mots, ou de formes fléchies de mots, extraits de textes divers. Elle affecte alors un même poids à chaque forme du lexique, puisque celui-ci ne contient qu'une seule occurence par forme.

L'étude du taux de couverture lexicale que nous avons effectuée sur le corpus de BREF, relève tout à fait de cette seconde approche.

2.1. LE CORPUS BREF

Les textes qui composent ce corpus proviennent du milieu journalistique. Les domaines abordés sont donc relativement vastes (finance, géographie, politique, culture, spectacle, ...).

2.1.1. ETAT DES DONNEES EXAMINEES

Les statistiques présentées ici, portent sur le lexique des mots, ou plutôt sur celui des formes fléchies figurant dans un corpus de textes constitué en vue de la création d'une base de données de parole enregistrée : la base de données BREF. Celle-ci est destinée à l'évaluation de systèmes de reconnaissance de grands vocabulaires. Elle est développée au LIMSI dans le cadre du GDR-PRC Communication Homme-Machine [Lamel, 91].

Nous nous intéresserons plus particulièrement à la composition du vocabulaire apparaissant dans les textes du corpus BREF. Celui-ci nous a été communiqué sous la forme d'une liste de 84 900 mots, liste que nous appellerons par la suite LexBref. Chaque mot est écrit en minuscules ; de plus les signes de ponctuations et autres signes non alphanumériques ont été effacés. La distinction entre nom propre et nom commun est donc complètement perdue, de même que les repérages typographiques conventionnels des sigles, des abréviations et de certains mots composés.

2.1.2. TAUX DE COUVERTURE LEXICALE

Nous avons procédé à la comparaison des formes de LexBref avec celles générées à partir de notre base de données lexicales BDLEX.

La version BDLEX-1 comporte 23 000 entrées et permet d'accéder à un corpus comptant environ 270 000 formes fléchies. L'extension de ce lexique à 50 000 entrées constitue la version BDLEX-2.

Dans la figure 1 nous avons représenté le pourcentage de formes de LexBref qui ont été trouvées dans BDLEX-1 et BDLEX-2.

La colonne (1) donne les résultats obtenus à partir d'une recherche lexicale directe. Nous avons ensuite supposé que d'autres formes pouvaient être trouvées, moyennant la correction d'une faute d'accent.

Les résultats portés en colonne (2) du tableau ont été obtenus en utilisant le correcteur orthographique et typographique VORTEX décrit dans [Pérennou, 86, 91], [Pécatte, 92].

Corpus de référence \ Nbre de formes trouvées	Recherche directe (1)	Fautes d'accent (2)	Pourcentage
BDLEX-1 (23 000 entrées)	40 931	1 542	50%
BDLEX-2 * (50 000 entrées)	9 415	183	11, 3%
LexBref par rapport à BDLEX-2	50 346	1725	**61,3%**

* : Complémentaire de BDLEX-1 par rapport à BDLEX-2

Fig.1 *- Résultats obtenus par comparaison de LexBref aux formes générées à partir de BDLEX-1 et BDLEX-2.*

Pour affiner l'analyse de LexBref, nous nous sommes intéressés aux sigles et aux abréviations qui pouvaient figurer dans ce corpus. Pour cela nous l'avons comparé à une liste de sigles, fournie par M. Plénat dans le cadre du GDR-PRC CHM, et à une liste d'abréviations. Les résultats de cette recherche sont portés dans la figure 2.

Corpus de référence	Formes trouvées	Pourcentage
Sigles de Plénat (1 000 sigles)	380	0,45%
Abréviations (280 abrév.)	70	0,08%
Sigles et abréviations de LexBref reconnues	450	**0,53%**

Fig.2 *- Résultats obtenus par comparaison de LexBref à une liste de sigles et une liste d'abréviations.*

En observant les figures 1 et 2 on constate qu'un ensemble important de formes de LexBref, environ 38%, n'ont pas été identifiées. Avant d'étudier plus précisément la structure de ce lexique résiduel, que nous appellerons désormais LexR, il est important de resituer les résultats obtenus par rapport au contexte général d'étude de la couverture lexicale.

2.2. RESULTATS CLASSIQUES

Pour les raisons que nous avons déjà évoquées, les résultats présentés dans le paragraphe précédent ne peuvent pas être directement comparés aux taux de couverture classiques, tels que ceux donnés en exemple ci-dessous.

L'étude de P. Guiraud [Guiraud, 59] établit que 100 mots bien choisis assurent un taux de couverture d'environ 60%, tandis que 1 000 mots couvrent 85% et 4 000 mots 97,5%. La couverture des 2,5% restant peut être assurée par un corpus de 40 000 mots.

D'autres études montrent qu'en terme de formes de mot, le taux de couverture est dépendant de la langue considérée. On peut citer notamment les statistiques établies à partir d'un corpus de lettres d'affaires, par Averbuch [Averbuch, 87] pour l'anglais et Mérialdo [Mérialdo, 88] pour le français, et desquelles il ressort que le taux de couverture assuré en anglais par un lexique de 20 000 formes (~97,5%) est équivalent à celui assuré, en français, par un lexique 10 fois plus important [Pérennou, 90].

Un dernier exemple est celui de l'enquête réalisée sur le *Français élémentaire* et présentée dans le Grand Larousse de la Langue Française [Guilbert, 71]. A la différence des études précédentes celle-ci porte sur le français *parlé*. Le corpus utilisé compte 312 135 mots (ou formes) qui correspondent en réalité à 7 995 vocables différents. Ces mots, classés selon leur fréquence d'apparition, ont permis de constater qu'un locuteur moyen connaissant les 38 mots les plus fréquents, peut identifier 50% des mots d'une conversation ; avec 278 mots ce pourcentage passe à 80%. Il faut noter toutefois, que les mots les plus fréquents sont pour l'essentiel des mots grammaticaux.

D'autres résultats statistiques de cette nature sont présentés dans [Catach, 84].

Ces taux de couverture, relativement élevés, sont obtenus selon la première approche, où la fréquence d'apparition des mots joue un rôle très important (cf. §2). Par conséquent, les formes rares ou très spécialisées, de poids généralement très faible, constituent une très petite partie des mots rejetés.

Ceci n'est plus vrai dès que l'on attribue un poids identique à chaque forme rencontrée. C'est pourquoi, dans un deuxième temps nous nous sommes intéressés au contenu du corpus résiduel LexR (cf. §2.1.2), ainsi qu'aux différentes causes de la non-identification d'une forme.

3. ANALYSE DES MOTS INATTENDUS

Pour effectuer cette deuxième phase d'analyse, nous avons fait intervenir plusieurs procédures non lexicales afin de mettre en évidence les différents types d'inattendus et leurs proportions respectives dans le corpus initial, LexBref. Nous avons ensuite étudié les possibilités de traiter ces mots, notamment en faisant appel à un ensemble de connaissances morphologiques.

3.1. NATURE DES INATTENDUS

♦ *Les formes numériques*
Le corpus LexBref comporte des nombres cardinaux et ordinaux, exprimés en chiffres arabes (*1991*, *200^{e}*,...), ou en chiffres romains (*XVII^{e}*, *XV*, ...). On trouve également des nombres traduisant un pourcentage (*5%*, *75%*, ...). Ces unités représentent environ 1,5% de ce corpus.

♦ *Les mots étrangers et les noms propres*
La grande diversité des sujets abordés dans un quotidien et la portée internationale des faits relatés font que de nombreux mots étrangers apparaissent dans les textes (*amnesty*, *congress*, *perestroïka*, *glasnost*...).

Une analyse basée sur des critères particuliers, comme l'étude des finales de mots n'appartenant pas à la langue française mais fréquentes dans d'autres langues, ou encore caractéristiques de noms propres (*-y*, *-ess*, *-ski*, *-nn*, *-ff*, *-v*, *-oux*, ...), nous a permis de distinguer, dans le corpus LexR, un premier groupe de noms propres (*Stravinski*, *Roscoff*, *Gorbatchev*, *Bonn*, *Châteauroux*...) et de mots d'origine étrangère, (*academy*, *press*, ...). Ceci correspond environ 15,5% du corpus initial.

♦ *Les néologismes*
La création lexicale est un phénomène linguistique fréquent dans les médias : *groupuscularisation*, *zapping*, ... Beaucoup de mots sont créés à partir de noms propres issus des milieux politique, artistique ou littéraire : *antigaulliste*, *mitterrandien*, *maccarthysme*, *hitchcockien*, *nabokovien*, ...

La plupart sont produits par dérivation mais il existe de nombreux exemples de mots obtenus par composition, comme celui de l'adjectif *vrai-faux* (*vraie-fausse facture*, *vrai-faux passeport*, ...). Quelques néologismes sont obtenus selon des procédés plus marginaux comme le verlan (*ripoux*, *chébran* ...) et les mots-valises (*motel*, *confipote* ...).

Nous avons examiné les formes correspondant à des néologismes dérivationnels, construits de manière régulière par application de processus dérivationnels sur un mot de la langue ou un nom propre.

A partir d'une liste d'affixes productifs comme les préfixes *anti-*, *dé-*, *inter-*, *néo-*, *sur-*, ... et les suffixes *-ation*, *-ien*, *-isme*, *-iste*, *-is(er)*, *-ité*, *-ment*, ..., nous avons procédé à une recherche dans LexR qui nous a permis d'estimer respectivement à 0,5% et 5,5% les mots de LexBref analysables comme préfixés ou bien suffixés —lors du traitement des suffixes nous avons pris en compte les variations flexionnelles (par exemple les mots comme *hitchcockiennes* sont détectés).

La figure 3 reprend les différentes estimations faites dans cette deuxième phase de traitement.

Critères de recherche	Exemples	Pourcentage par rapport à LexBref
Nombres	*1991, XXVIIe,...*	1,5%
Mots étrangers et noms propres	*congress, amnesty, roscoff, gorbatchev*	15,5%
Mots supposés préfixés	*interafricain, néobaroque,...*	0,5%
Mots supposés suffixés	*hitchcockiennes, groupuscularisation, zapping, ...*	5,5%
Mots extraits de LexBref par procédure non lexicale		23%

Fig.3- Analyse du corpus résiduel LexR.

Parmi les 15% restant, on trouve notamment des sigles qui n'ont pas été répertoriés dans la liste de référence que nous avions à notre disposition (TF1, ADN,...). On trouve encore des néologismes, des noms propres et des mots étrangers d'emprunt, pour lesquels aucune procédure non lexicale n'a pu être appliquée. Enfin, on rencontre des mots incorrectement écrits, le plus souvent à la suite d'une faute typographique, et d'autres qui seraient reconnus par un lexique plus étendu que BDLEX-2.

Pour effectuer cette analyse, nous avons mis en œuvre un ensemble de procédures non lexicales, c'est-à-dire qui n'effectuent aucune consultation de lexique. On constate que, parmi les mots identifiables selon cette méthode, seuls les néologismes obtenus par dérivation peuvent faire l'objet de recherches plus approfondies.

3.2. ETUDE MORPHOLOGIQUE

Les mots classés comme mots étrangers ou noms propres, ne peuvent être identifiés comme tels que grâce à l'utilisation d'un lexique approprié (lexique de mots étrangers ou lexique de noms propres), ou encore grâce à la connaissance apportée par une personne moyennement cultivée, capable de juger de l'acceptabilité des mots.

En revanche, l'étude de la structure morphologique des mots dérivés révèle que, dans de nombreux cas, ces mots peuvent être rattachés à une entrée connue du lexique. En effet, les mots dérivés sont définis comme construits par association d'un élément lexical, appelé généralement la *base*, et d'un élément affixal.

3.2.1. REGLES DERIVATIONNELLES

L'utilisation de règles traduisant le fonctionnement de processus dérivationnels réguliers doit permettre de procéder à l'analyse formelle d'un mot construit. Chaque règle est assimilable à un opérateur dérivationnel (OD) dont la structure générale (cf. Fig.4) est la suivante [Corbin, 86] :

$$OD = RC + OM + OS$$

♦ Le premier constituant (RC) représente le rapport catégoriel qui existe entre la catégorie syntaxique de la base et celle du mot qui en est dérivé.

♦ Le deuxième composant (OM) correspond à l'opération qui fournit la structure morphologique du mot dérivé. Elle est différente selon qu'il s'agit d'une préfixation et d'une suffixation.

♦ Enfin, l'opération sémantique (OS) permet d'attribuer un sens morphologique au mot dérivé. Cette information est obtenue par composition (*) du sens (S) associé à l'affixe avec celui de la base. Il faut cependant préciser que le, ou les sens attribués par l'usage à un mot dérivé peuvent être différents de celui obtenu par application de l'opération sémantique.

RC : $CS_{Base} \rightarrow CS_{Dérivé}$

OM (Préf) : Dérivé = Partie_préfixale + Base
(Suff) : Dérivé = Base + Partie_suffixale

OS : S(Dérivé) = S(Affixe)*S(Base)

Fig.4 - Composants d'un opérateur dérivationnel.

Bien que ces opérateurs aient été initialement conçus pour construire des mots dérivés à partir d'un affixe et d'une base, nous pouvons les utiliser pour retouver la représentation morphologique de la base connaissant celles du mot dérivé et de l'affixe —[Ferrané, 91] pour le traitement morphologique dans BDLEX.

Lorsqu'on applique un processus dérivationnel, on met en présence deux éléments, la base et l'affixe. De manière générale la forme résultante et plus particulièrement la forme de l'élément de gauche est conditionnée par l'élément de droite.

Dans le cas d'une préfixation, la forme prise par l'affixe dépend du contexte introduit par la base. Ainsi, ce sont les variantes *ir-*, *il-*, *im-* du préfixe *in-* qui s'appliquent respectivement en contexte "r", "l" et "b,m,p" (*réel / irréel, logique / illogique, possible / impossible, ...*).

Lors de la formation d'un mot suffixé, la forme prise par la base peut être fonction de l'affixe appliqué. Ainsi, les mots suffixés *direction* et *dirigeable* ont une base commune (*diriger*), mais les radicaux utilisés sont différents : *direct-* pour la suffixation en *-ion* et *dirig-* pour celle en *-able*. Si on applique le suffixe *-ité* à un mot déjà suffixé par *-able*, c'est la variante *-abil-* qui est employée (acceptable / acceptabilité).

L'analyse morphologique dérivationnelle pose donc le problème de la prise en compte des variations morpho-phonologiques affixales et des radicaux.

Un opérateur dérivationnel relatif à un affixe peut être muni d'une liste répertoriant le plus exhaustivement possible, les différentes variantes admises par l'affixe, ainsi que le contexte associé, lorsque celui-ci est connu.

Le second aspect du problème peut être résolu si on dispose d'un lexique régroupant les radicaux, ou bases non autonomes, fréquemment rencontrés en morphologie. Ce lexique pourrait par exemple contenir les radicaux *vict-* et *vinc-* associées à l'entrée *vaincre* et que l'on retrouve dans les mots dérivés *victoire* et *invincible*.

En mettant sous forme de règles les processus dérivationnels réguliers les plus productifs, on peut être en mesure de reconstituer les associations *base/affixe* supposées à l'origine de nombreux mots dérivés inattendus rencontrés dans les textes. La solution proposée est alors validée par recherche lexicale sur la base. En effet, si la base est connue du lexique et si les informations qui lui sont associées sont compatibles avec celles de la règle dérivationnelle utilisée alors le mot dérivé étudié peut être considéré comme identifié de manière indirecte.

Cette démarche a été utilisée pour compléter l'analyse lexicale de LexBref, et étudier plus particulièrement les mots supposés affixés qui ont pu être mis en évidence dans LexR (cf. §3.1).

3.2.2. RESULTATS OBTENUS SUR LEXR

Nous avons associé chaque mot à traiter avec sa base supposée, en faisant abstraction du problème des radicaux évoqué précédemment.

La base d'un mot préfixé est obtenue en enlevant la partie affixale (*incorrect - in- = correct*). Pour un mot suffixé, la transformation a effectuer est plus complexe. En effet, en retranchant la partie suffixale, on obtient un élément lexical qui peut ne pas correspondre directement à la base. Dans de nombreux cas, l'ajout d'une chaîne terminale, comme une marque d'infinitif par exemple, résout le problème. La transformation effectivement réalisée pour obtenir la base consiste alors à substituer la partie suffixale par la terminaison associée : *able/-er* : *acceptable / accepter*.

Ce type d'association, *partie_suffixale / terminaison*, est souvent difficile à déterminer. On constate cependant que l'étude de la partie suffixale étendue à son contexte gauche, facilite considérablement la définition de telles associations. La figure 5 présente plusieurs associations, *partie_suffixale_large / terminaison*, à prendre en compte dans la règle dérivationnelle associée au suffixe *-ment*, formateur d'adverbe à partir d'une base adjectivale.

P_Suff_Large	Terminaison	Exemple
-ablement	-able	*considérablement*
-alement	-al	*généralement*
-ellement	-el	*usuellement*
-ivement	-if	*tardivement*
-emment	-ent	*prudemment*
-amment	-ant	*savamment*
-ément	-e	*commodément*
-ment	-	*joliment*

***Fig.5** - Associations partie_suffixale_large-terminaison pour la formation d'adverbes en -ment.*

Si pour certains suffixes (*-ment, -ation, -age, ...*) mettre en évidence ces associations est relativement simple, pour d'autres, cela pose problème (*-ien, -eux, ...*).C'est pourquoi, l'opération de reconstitution de la base n'a été effectuée que pour une partie seulement des suffixes pris en compte lors de la sélection des mots de LexR supposés suffixés.

La figure 6 présente donc les premiers résultats obtenus après l'étude de la structure morphologique d'une partie des mots affixés de LexR, et après la recherche lexicale effectuée sur les bases obtenues.

Mots de LexR supposés préfixés	Traités	818	
	Base connue de BDLEX	496	→ 60,5% des mots préfixés traités
Mots de LexR supposés suffixés	Traités	2051	
	Base connue de BDLEX	1252	→ 61% des mots suffixés traités

***Fig. 6** - Résultats obtenus après étude de la structure morphologique d'un sous-ensemble de mots affixés.*

Dans les deux cas, plus de la moitié des bases traitées, 60% environ, ont été trouvées dans BDLEX. L'ensemble des bases non identifiées est composé de mots, eux-mêmes dérivés et inconnus de BDLEX (1), ou bien de formes comportant une ou plusieurs fautes typographiques (2). Dans ce cas il y a répercussion de la faute orthographique ou typographique du mot dérivé sur la base qui lui a été associée.

(1) *groupuscularisation = groupuscularis(-er) + -ation*
(2)**goudonnage = *goudonn(-er) + -age*

Le problème des mots inconnus et de l'étude de la structure morphologique des mots a déjà été abordé dans le cadre de la correction automatique, notamment dans [Fournier, 87] et [Sabah, 88]. En effet, l'utilisation des connaissances morphologiques peut être considérée comme une méthode complémentaire de la correction, permettant de poursuivre l'analyse des textes lorsque celle-ci pourrait être bloquée.

Les mots candidats, proposés par un système de correction en remplacement d'un mot erroné, sont parfois trop éloignés de l'entrée lexicale à laquelle le mot se rapporte réellement.

Prenons deux exemples de formes de LexR, l'une erronée, l'autre correcte sur le plan orthographique : **investisement* et *aboutissement*.

L'application du correcteur VORTEX, fonctionnant sur la version BDLEX-2 du lexique, donne respectivement comme solutions : *investissement* et *abrutissement*. Dans le second cas la forme proposée ne doit pas être acceptée.

L'utilisation d'une méthode basée sur l'étude de la morphologie du mot permet de rattacher le mot correct *aboutissement* au mot *aboutir* connu de BDLEX, et ce,

en le considérant comme dérivé de la base verbale *aboutir* par application du suffixe *-ment* formateur de noms masculins.

Sur le plan lexical, cette méthode apporte un complément d'informations non négligeable pour le traitement d'une partie des mots inconnus rencontrés lors de l'analyse d'un corpus, tel que LexBref. Dans une perspective plus large qui est celle du traitement des textes réels, cette méthode, et en particulier l'utilisation d'opérateurs dérivationels, peut également donner accès à des informations d'ordre syntaxique, voire d'ordre sémantique, utiles pour la compréhension des textes traités (cf. §3.2.1).

4. CONCLUSION

Comme nous l'avons illustré à partir du lexique extrait du corpus de BREF, le traitement automatique de la parole et des textes requiert un ensemble de matériaux lexicaux importants et variés, incluant les sigles et les abréviations, ainsi que des éléments de morphologie tels que les radicaux. Ces matériaux doivent être complétés par divers outils linguisticiels améliorant le traitement (correction, analyse morphologique, ...).

Pour cela, il faut non seulement prendre en compte les besoins classiques aux plans morphologique et syntaxique, mais encore ceux, plus particuliers, relatifs aux inattendus variés qui apparaissent dans les textes et les messages vocaux.

Le projet BDLEX développé dans le cadre du GDR-PRC Communication Homme-Machine, avec comme objectif de rendre disponibles des matériaux et des outils linguistiques répond partiellement à ces besoins.

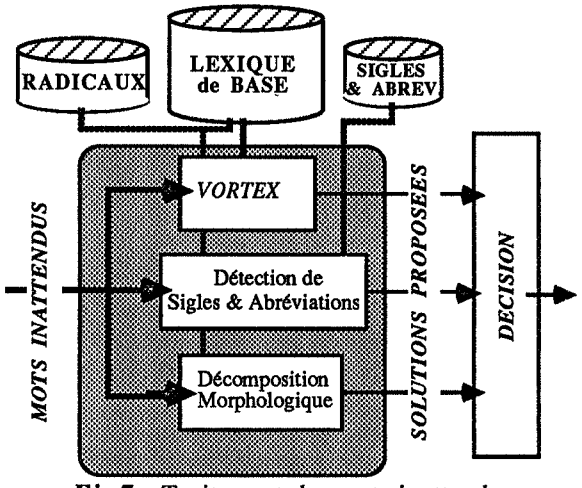


Fig.7 - Traitement des mots inattendus

L'analyse que nous avons effectuée met en évidence l'insuffisance d'un lexique de 50 000 mots et renforce l'idée que la prise en compte de textes réels doit passer par un système combinant correction et analyse de la structure morphologique des mots, comme cela est illustré dans la figure 7. Cette étude mérite d'être complétée par des résultats prenant réellement en compte la fréquence d'apparition des mots dans les textes.

Les extensions en cours du projet BDLEX visent entre autres, à l'enrichissement du vocabulaire et au développement des traitements morphologiques.

5. BIBLIOGRAPHIE

[Averbuch, 87] A. Averbuch et 21 co-auteurs, *Experiment with the TANGORA 20,000 Word Speech Recognizer*, CH2396-0/37/0000-0701, 1987.

[Catach, 84] N. Catach, *Les listes orthographiques de base du français (LOB) - Les mots les plus fréquents et leurs formes les plus fréquentes*, Nathan Recherche, 1984.

[Corbin, 86] D. Corbin, *Morphologie dérivationnelle et structuration du lexique*, Thèse de doctorat d'état de l'Université Paris VIII, 1986.

[Ferrané, 91] I. Ferrané, *Base de données et de connaissances lexicales morphosyntaxiques*, Thèse de doctorat de l'Université Paul Sabatier, Toulouse III, septembre 1991.

[Fournier, 87] J.P. Fournier 1 coll., *Traitement des mots inconnus dans un système de questions-réponses en langue naturelle*, Reconnaissance des Formes et Intelligence Artificielle, 6ème congrès, AFCET, Antibes, 16-20 Novembre 1987, vol II, pp. 653-667.

[Guilbert, 71] P. Guilbert, *De la notion de vocabulaire essentiel - Introduction au Grand Larousse de la Langue Française*, Tome 1, p. LXXXII, Larousse, Paris, 1971.

[Guiraud, 59] P. Guiraud, *Problèmes et méthodes de la statistique linguistique*, D. Reidel Pub. Company, 1959.

[Lamel, 91] L.F. Lamel & coll., *BREF, a Large Vocabulary Spoken Corpus for French*, Proceedings of EUROSPEECH 91, Genova, 24-26 September 1991, Vol.2, pp. 505-508.

[Mérialdo, 88] B. Mérialdo, *Multi-Level Decoding for Very Large Size Dictionary Speech Recognition*, IBM Journal of R&D, 1988.

[Pécatte, 92] J.M. Pécatte, *Tolérance aux fautes dans les interfaces homme-machine*, Thèse de doctorat de l'Université Paul Sabatier, Toulouse III, janvier 1992.

[Pérennou, 86] G. Pérennou, *La vérification et la correction automatique des textes : le système VORTEX*, Technique et Science Informatique, n°4, 1986, pp. 285-305.

[Pérennou, 90] G. Pérennou, *Le projet BDLEX de base de données et de connaissances lexicales et phonologiques*, Société de Neuropsychologie de Langue Française, (éd. J.L. Nespoulous & M.Leclercq), Paris, 1990, pp. 117-140.

[Pérennou, 91] G. Pérennou & coll., *Composantes phonologique et orthographique de BDLEX*, Deuxièmes journées du GDR-PRC Communication Homme-Machine, EC2 Editeur, Toulouse, 29-30 Janvier 1991, pp. 351-362.

[Pérennou, 92] G. Pérennou & coll., *Le projet BDLEX de base de données lexicales du français écrit et parlé*, Séminaire Lexique, Toulouse, 21-22 Janvier 1992.

[Sabah, 89] G. Sabah, *L'intelligence artificielle et le langage naturel - Processus de compréhension*, Vol 2, Hermès, Paris, 1989.

19es J.E.P. - Bruxelles, 19 au 22 mai 1992.

SEGMENTATION EN EVENEMENTS PHONETIQUES ET MODELES MARKOVIENS HMM POUR L'ETIQUETAGE PHONOTYPIQUE

A. FARHAT, G. PERENNOU, N. VIGOUROUX

IRIT - URA CNRS 1399 - UNIVERSITE PAUL SABATIER
118, route de Narbonne - 31062 Toulouse Cedex

Résumé

Après avoir posé le problème de la segmentation automatique de la parole et donné les principales approches existant, nous détaillons les différentes composantes de notre système de segmentation automatique. Ce système fonctionne en 2 passes: on effectue dans un premier temps la segmentation du corpus de la parole en événements phonétiques par le biais du système SAPHO, et dans une seconde étape, en utilisant ces événements phonétiques obtenus, on procède à l'alignement automatique des chaînes phonétiques par l'algorithme de Viterbi. Nous avons regroupé les différentes unités phonétiques en 10 classes majeures, chaque classe étant modélisée par un modèle Markovien HMM discret. L'expérimentation s'est faite en 4 étapes: unités indépendantes de contexte, unités dépendantes de contexte gauche, unités dépendantes de contexte droit et unités dépendantes de contexte gauche et droit. Enfin nous donnons, pour chacune des expériences citées, quelques résultats obtenus sur le corpus d'EUROM0.

1. INTRODUCTION

Dans le domaine du traitement automatique de la parole, on utilise de plus en plus souvent des corpus enregistrés. Depuis quelques années se développent plusieurs projets de bases de données de sons. On peut citer par exemple DARPA TIMIT [NIST 88], BDSONS du PRC-GRECO Communication Homme Machine [Carré 84], EUROM0 de SAM [SAM Report 89] et la base ATR du Japon [Takeda 87].

Dans le cadre de certaines applications de la parole (apprentissage, évaluation, ...) la segmentation et l'étiquetage (en unités phonétiques ou autres) de ces corpus peut s'avérer nécessaire. Ces tâches effectuées par l'expert phonéticien sont considérées comme longues et fastidieuses. D'où l'intérêt d'étiqueter automatiquement les corpus de parole.

Dans cette communication, après avoir exposé les différentes approches d'étiquetage automatique déjà développées, nous tentons de situer notre système d'alignement dans le paragraphe 2. Le paragraphe 3 est consacré à l'architecture de notre système avec une description succincte de ses différentes composantes, en particulier le système de segmentation en événements phonétiques choisi. Dans le paragraphe 4, nous décrivons les modèles HMM choisis et le processus d'apprentissage de modèles appliqué. Le paragraphe 5 décrit l'alignement effectué par l'algorithme de Viterbi avec les résultats obtenus au cours de différentes expériences. Enfin nous donnerons nos conclusions et nos perspectives pour la suite des travaux.

2. TOUR D'HORIZON DES SYSTEMES D'ETIQUETAGE AUTOMATIQUE

A ce jour différentes méthodes d'étiquetage automatique ont été développées. On peut citer entre autres l'algorithme de comparaison dynamique [Chamberlain 83], [Wagner 81], ou des méthodes à base de modèles Markoviens cachés (notés HMM) [Baker 74]. D'autres approches nécessitent d'obtenir, dans un premier temps les événements phonétiques (ou plus généralement des segments phonétiques) qui seront par la suite utilisés pour l'alignement des chaînes d'unités phonétiques.

Parmi les différentes approches de segmentation en événements phonétiques on peut noter les méthodes à base de règles [Dixon 76], [Dalsgaard 87], les méthodes basées sur modèles de rupture [André-Obrecht 88] et les approches connexionnistes [Glass 88], [Dalsgaard 89].

Pour notre part, nous avons opté pour une segmentation préalable de la parole en événements phonétiques par une méthode mixte, basée à la fois sur l'étiquetage des trames et sur les discontinuités majeures effectuée par le système SAPHO [Pérennou 91]. Nous avons ensuite modélisé les différentes unités de la notation phonétique par des modèles markoviens cachés (noté HMM) discrets, et finalement nous avons utilisé

l'algorithme de Viterbi [Forny 73] pour procéder à l'alignement de la notation phonétique avec les segments de SAPHO.

Les corpus de parole utilisés sont les corpus de BDSONS et d'EUROM0, composés de chiffres isolés et de parole continue. Ces corpus ont été étiquetés manuellement. On pourra se reporter à [Autesserre 89] pour une description détaillée.

3. ARCHITECTURE DU SYSTEME

Notre système d'alignement est composé de cinq modules qui peuvent être considérés comme des sous-systèmes indépendants.

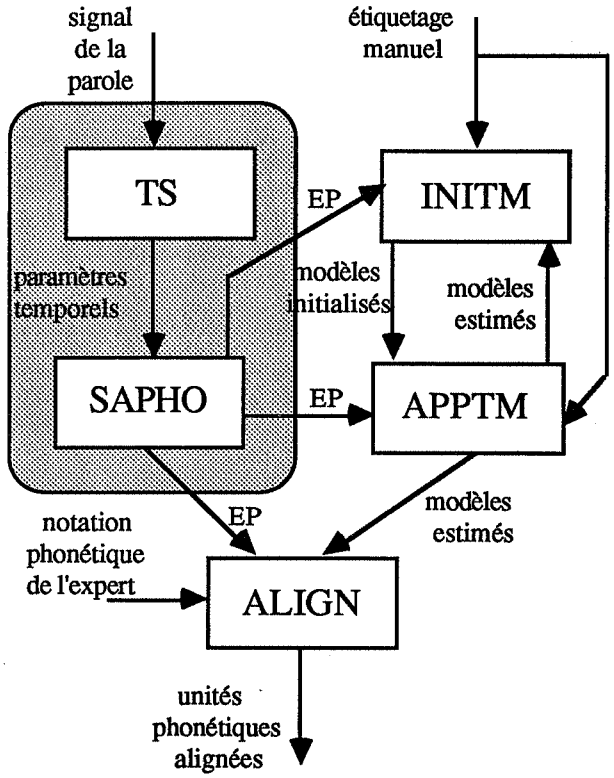


Figure 1. Schéma de liaison entre les différents modules du système d'alignement.

Ces modules sont :

- Le processeur de traitement du signal (noté TS). Il calcule les paramètres temporels nécessaire à la segmentation en événements phonétiques.
- Le module de segmentation et d'étiquetage de la parole en événements phonétiques (noté SAPHO).
- Le module de définition et d'initialisation automatique de la structure topologique des modèles HMM (correspondant aux différentes unités phonétiques de la notation) avec un ensemble restreint de données étiquetés manuellement (noté INITM).
- Le module d'apprentissage des modèles HMM par estimation/ réestimation utilisant l'algorithme de Viterbi sur un sous-ensemble du corpus de parole étiqueté manuellement (noté APPTM).
- Le module d'alignement automatique des unités phonétiques sur les événements phonétique en appliquant l'algorithme de Viterbi (noté ALIGN).

Nous avons effectué une évaluation des résultats de l'alignement en comparaison avec l'étiquetage manuel au moyen des méthodes d'évaluation développées dans le cadre du projet SAM-ESPRIT.

Les deux parties TS et SAPHO ont été développées précédemment à l'IRIT département CHM, pour de plus amples informations se référer à [Pérennou 91]. Dans le cadre de notre communication on rappellera les points suivants sur les événements phonétiques.

Chaque événement phonétique (noté EP) est défini par :

- une étiquette,
- un indice de durée,
- un indice syllabique.

Les étiquettes des événements correspondent aux classes phonétiques majeures (voir table I).

L'indice de la durée peut prendre les valeurs 1, 2, 3.

- 1 : EP dont la durée est inférieure à 4 ms.
- 2 : EP dont la durée est comprise entre 4 et 8 ms.
- 3 : EP dont la durée est supérieure à 8 ms.

L'indice syllabique peut prendre les valeurs 0, 1, 2, 3.

- 0 : correspond à la convexité positive de l'amplitude qui est représenté par une cuvette sur la courbe,
- 1 : correspond à l'attaque syllabique,
- 2 : correspond au noyau syllabique,
- 3 : correspond à la fin ou le coda syllabique.

Ce que l'on appelera dans ce qui suit événement phonétique (EP) sera un élément e_{ds} où :

$e \in \{ K, \ldots, S \}$
$d \in \{ 1, 2, 3 \}$
$s \in \{ 0, 1, 2, 3 \}$

Code	Signification
K	voyelle
N	voyelle aigu
M	voyelle grave
L	vocalique aigu
U	vocalique grave
O	occlusif voisé
Q	occlusif sourd
X	fricatif de type R
F	fricatif faible
Z	fricatif voisé
S	fricatif sourd

Table I. Les événements phonétiques

4. STRUCTURE DES MODELES HMM

4.1 Description des unités de base

Nous avons réparti les différentes unités phonétiques en dix classes majeures. Chacune de ces classes sera représentée par un modèle HMM discret. Ces classes sont les suivantes:

- V : voyelle forte : a ɑ ɛ o ø ɔ æ œ Œ
- I : voyelle faible : i ə y u
- M: voyelle nasale : ɔ̃ ɑ̃ œ̃ ɛ̃
- L : liquide et nasale : l m n ɲ
- R : regroupe les 2 phonème: r v
- S : fricative sourde : f s ʃ
- Z : fricative voisée : z ʒ
- Q : occlusive sourde : p t k
- O : occlusive voisée : b d g
- U : semi-voyelle : j w ɥ

Dans un premier temps nous avons procédé à une expérimentation avec les modèles de classe de phonème indépendants du contexte. Ensuite nous avons procédé à différentes expériences prenant en compte successivement:

- le contexte gauche,
- le contexte droit,
- le contexte gauche et droit. Ces modèles seront appelés contextuels.

Pour limiter le nombre de modèles contextuels, nous avons réuni les dix classes de phonèmes en quatre grandes classes de contexte:

- FR : fricative (sourd et voisé): S et Z
- OC : occlusive sourd : Q et pause
- VC : liquide, nasale et occlusive voisé: R,L et O
- VY : voyelles et semi-voyelle: V, I, M et U.

La définition et l'initialisation des modèles HMM associés à ces classes fera l'objet du paragraphe suivant.

4.2 Initialisation de modèles HMM

Les modèles HMM définis sont discrets. Nous avons opté pour deux structures de modèle HMM (figure 2 et 3):

Les états 0 et F sont des états fonctionnels: sans émission d'événement phonétique (figure 2 et 3). A chaque état i : 1...10 (respectivement 1...7) est associé une loi de probabilité d'émission d'événements phonétiques e_{ds}.

Le choix structurel du modèle HMM I a été effectué pour permettre la différenciation des lois de probabilité d'émission des réalisations à 1 événement phonétique, à 2 EP, à 3 EP et à 4EP ou plus. En effet nous avons remarqué une grande disparité des lois de probabilité dans ces différentes réalisations. Cependant pour les réalisations à plus de 4 EP, le noyau de l'unité phonétique se comportait, dans une majorité de ces réalisations, de manière similaire.

Au cours de nos expérimentations nous avons également observé que certaines unités phonétiques (les voyelles nasales par exemple) se réalisaient toujours à plus de 3 événements phonétiques. Cette observation nous a conduit à choisir le modèle HMM II pour modéliser ces classes d'unités phonétiques. Dans le cadre des expériences présentées dans cette communication, nous avons utilisé le modèle HMM II pour modéliser les voyelles nasales (classe M). Le modèle HMM I est utilisé pour toutes les autres classes de phonèmes.

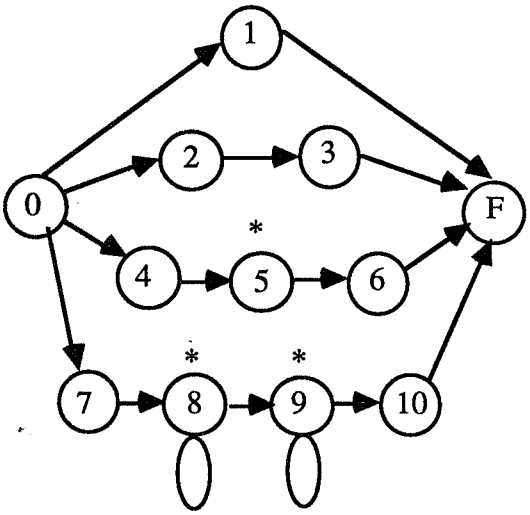


Figure 2. Modèle HMM I à 10 états émetteurs

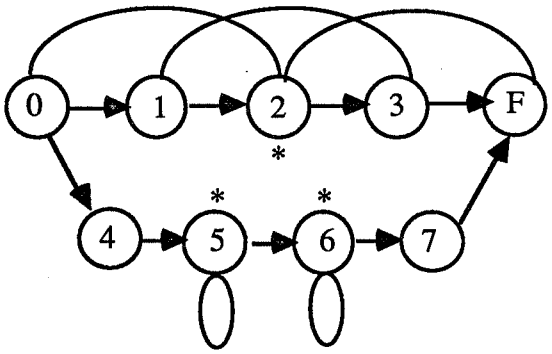


Figure 3. Modèle HMM II à 7 états émetteurs

4.3 Hypothèses restrictives

Dans le cas de modèles contextuels nous avons un total de 160 modèles dont 144 seront de la forme I et 16 de la forme II. Le nombre important (1552) de lois de probabilité d'émission à estimer nous a amené à faire un certain nombre d'hypothèses restrictives qui sont détaillées dans ce qui suit.

D'une part nous avons défini la notion d'état stable: ce sont les états pour lesquels la loi de probabilité d'émission sera la même quel que soit le contexte gauche et droit. Ces états sont marqués d'une * sur les figures 2 et 3. On pourrait les considérer comme représentant le noyau de l'unité phonétique quand celle-ci se réalise en 3 événements phonétiques ou plus. Avec cette hypothèse restrictive nous avons réduit le

nombre de loi de probabilité d'émission à 1102. Ce qui reste tout de même assez élevé.

Nous avons donc été confronté au problème de réunir suffisamment de données pour effectuer un bon apprentissage (Rappelons que chaque loi de probabilité est un vecteur de 132 éléments, certaines de ces lois ayant un petit nombre de réalisations dans le corpus d'apprentissage, la procédure d'estimation-réestimation serait pénalisante dans leur cas).

Pour remédier à ces problèmes nous avons choisi une politique d'apprentissage des modèles contextuels à plusieurs passes qui est détaillée dans le paragraphe suivant. Nous avons par ailleurs introduit une dernière étape de lissage des modèles estimés avant de procéder à l'alignement par l'algorithme de Viterbi. Ce lissage est du même type que celui effectué dans [Lee 89] basé sur le calcul d'une matrice de co-occurrence d'événements phonétiques.

4.4 Procédure d'apprentissage

Pour réaliser l'apprentissage des modèles HMM nous disposions de corpus de BDSONS (chiffres isolés) et d'EUROM0 (parole continue) étiquetés manuellement. Cet apprentissage est réalisé par la procédure d'estimation-réestimation appliquée aux trajectoires internes des modèles d'unités phonétiques, résultant de l'algorithme de Viterbi.

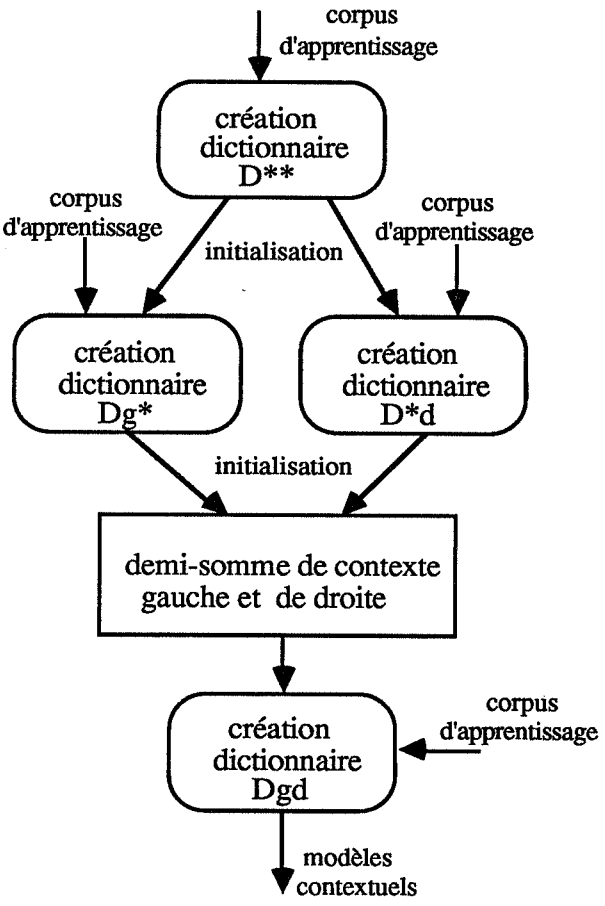


Figure 4. Schéma du processus d'apprentissage.

Pour pallier le problème du nombre insuffisant de réalisations des unités contextuelles dans le corpus d'apprentissage, nous avons procédé de la manière suivante :

- Dans un premier temps nous avons estimé les modèles HMM correspondant à nos classes de phonèmes indépendamment du contexte et obtenu un premier dictionnaire de modèles noté D**.
- Ces modèles indépendants du contexte obtenus lors de la première passe ont été utilisés pour initialiser les modèles dépendants de contexte gauche et dépendants de contexte droit. Puis nous avons ré-estimés ces 2 dictionnaires de modèles, que l'on notera respectivement Dg* et D*d dans la 2ème et la 3ème passe.
- Les modèles d'unités contextuelles ont été initialisés avec la demi-somme des modèles de la même unité dépendant de contexte gauche et dépendant de contexte droit. Dans une 4ème passe nous avons ré-estimé ces modèles contextuels et obtenu le dictionnaire Dgd.

Enfin, nous avons fixé un seuil minimum pour le nombre de réalisations dans le corpus d'apprentissage concourant à l'apprentissage d'un modèle HMM, pendant les phases 2, 3 et 4. Si pour une classe phonétique donnée ce seuil n'est pas atteint, le modèle ne sera pas estimé et sera remplacé par son modèle initial.

5. RESULTATS

Nous présentons les résultats obtenus lors de nos expérimentations sur le corpus d'EUROM0, en séparant le cas des modèles HMM lissés et non lissés (table II et III) et nous donnons pour chaque cas le taux d'erreurs d'alignement observés en comparaison avec un étiquetage manuel pour une précision de frontière de 20 ms.

%	erreurs SAPHO	Indépend CTX	CTX gauche	CTX droit	contextuel
données apprent.	5.4	12.9	12.9	9.7	6.8
données test	6.2	20.5	20.5	20.0	22.4

Table II . Taux d'erreurs obtenus pour les modèles non lissés

%	erreurs SAPHO	Indépend. CTX	CTX gauche	CTX droit	contextuel
données apprent.	5.4	12.9	12.6	9.4	7.7
données test	6.2	19.5	19.6	17.7	19.5

Table III . Taux d'erreurs obtenus pour les modèles lissés

En observant ces résultats, nous pouvons relever les points suivants :

• Les frontières omises par SAPHO ne sont pas rattrapées lors de l'alignement puisque celui-ci consiste à regrouper un ou plusieurs événements phonétiques. Cela représente environ 6% d'erreurs (soit 1/3 du nombre total d'erreur). Dans beaucoup de cas il s'agit de réalisations sous forme de "cluster" où une frontière objective n'est pas visible.

• Avec le lissage des modèles nous diminuons le taux d'erreurs de 1 à 3 % en moyenne sur les données de tests, quel que soit le contexte, sans dégrader les résultats obtenus sur les données d'apprentissage.

• Le contexte gauche ne semble rien apporter au résultat de l'alignement: nous avons obtenu les mêmes résultats qu'indépendamment de contexte.

• Avec le contexte droit nous améliorons sensiblement l'alignement effectué (0.5 à 3 %) autant pour les données de tests que pour les données d'apprentissage ([Lee 89] avait fait la même constatation dans le cadre de la reconnaissance phonétique). Nous avons remarqué que la grande partie des erreurs relevés pour ces modèles se situe dans le cas du contexte OC (occlusive sourde et silence). Ceci est dû au fait que l'on a appris dans le même modèle des unités suivies par un silence ou une occlusive sourde. Ce qui a entraîné une ambiguïté des lois de probabilités d'émission pour la partie finale de la réalisation des classes de phonèmes.

• Avec la combinaison des 2 contexte gauche et droit nous améliorons les performances pour les données d'apprentissage (5 à 6% par rapport aux modèles indépendant de contexte) tandis que l'on observe une dégradation importante sur les données de test (jusqu'à 2%). Les modèles ainsi obtenus dépendent fortement de leurs réalisations dans le corpus d'apprentissage. Pour résoudre ce problème il faut disposer d'un plus grand corpus d'apprentissage où pour chaque classe phonétique on aurait un nombre raisonnable de réalisations dans tous les contextes.

6. PERSPECTIVES ET CONCLUSIONS

Au stade actuel des expérimentations on obtient des résultats qui paraissent satisfaisants par rapport à ceux obtenus dans le cadre du projet SAM [SAM Report 89]. Celles-ci constituent notre premier train d'expériences qui nous permettent d'ores et déjà de mettre en évidence la nécessité d'avoir un corpus d'apprentissage assez important afin de bien modéliser le contexte. D'autre part l'analyse des erreurs observées a soulevé le problème de l'ambiguïté induite par le regroupement du silence et de l'occlusive dans la même classe de contexte, essentiellement pour les modèles dépendants de contexte droit.

Pour la suite de nos expériences nous envisageons de nous limiter au contexte droit, ce contexte étant celui qui donne les meilleurs résultats sur les données de tests tout en donnant de bons résultats sur les données d'apprentissage.

Jusqu'à présent, pour limiter le nombre déjà important de modèles bi-contextuels, nous avons réduit le nombre de classes de contexte à quatre. Les erreurs observées nous conduisent à penser que ce nombre est trop faible (en particulier il faut distinguer entre occlusive sourde et silence). En se limitant aux modèles dépendants de contexte droit, nous pouvons raffiner ces classes de contexte.

Nous envisageons aussi d'assouplir l'alignement pour tenir compte des clusters. Pour cela nous introduirons dans certains cas des arcs 0->F (c.f. figure 2 et figure 3) permettant de sauter une unité.

Enfin, la segmentation de SAPHO étant un système indépendant des corpus, des locuteurs et des langues, pour pouvoir utiliser notre méthode de segmentation dans une autre langue, il suffirait de définir:

• les différentes classes de phonèmes,
• les différentes classes de contexte
• et de disposer d'un corpus suffisamment représentatif de la langue pour effectuer l'apprentissage des modèles.

BILBIOGRAPHIE

[André-Obrecht 88] R.André-Obrecht, "A new statistical approach for the automatic segmentation of continuous speech signals" IEEE, Trans on ASSP, vol 36- n° 1, january 1988.

[Autesserre 89] D. Autesserre, G. Pérennou, M. Rossi, "Methodology for the transcription and labelling of a speech corpus", Journal of the International Phonetic Association, 1989, 19(1), pp. 2-15.

[Baker 74] J.K. Baker, "Machine-Aided Labelling of Connected Speech", in Working papers in Speech Recognition II, Computer Science Department

[Carré 84] R. Carré et al., "The French Language Database: defining, Planning and Recording Large Database" in Proceedings of ICASSP-IEEE, paper 42.10.1 San Diégo.

[Chamberlain 83] R.M. Chamberlain, J.S. Bridle, "ZIP: A Dynamic Algorithm for Time-Aligning Two Indefinitely Long Sequences", in Proceedings ICASSP-IEEE, April 1983, pp 816-9.

[Dalsgaard 87] P. Dalsgaard, A. Boekgaard, P. Holtse, "The HEAD System and its Approach to rules Based Acoustic-Phonetic Recognition of speech", Denmark, European Speech Technology, Volume 2, pp 5-8.

[Dalsgaard 89] P. Dalsgaard, "Semi-Automatic Labelling of Speech Data using a self organising neural network", in Proceedings of EUROSPEECH 1989, Paris, Vol.2, pp 541-544.

[Forny 73] G.D. Forny, "The Viterbi Algorithm", IEEE Mars 1973.

[Glass 88] J.R. Glass, V.W. Zue, "Multi-Level Segmentation of Continuous Speech", in Proceedings of ICASSP-IEEE, 1988,paper S 10.6.
[[Lee 89] K.F. Lee, H.W. Hon, "Speaker-Independent Phone Recognition Using Hidden Markov Models", IEEE Trans. on ASSP, vol 37, n° 11, Nov 1989.
[NIST 88] "An acoustic Phonetic Continuous Speech Database" in DARPA TIMIT CD-ROM, Prototype Version 1988.
[Pérennou 91] G.Pérennou, H. Kabré, N. Vigouroux, "Automatic SAPHO Segmentation of EUROM0, Multi-Lingual Speech Corpora Into Phonetics Events, SAM-Report, June 1991.
[SAM-Report 89] Rapport de la phase d'extension du projet SAM: Multi-Lingual Speech Input/output: Assessment, Methodology and Standardisation.
[Takeda 87] K. Takeda, Y. Sagisaka, S. Katagiri, "Acoustic phonetic labels Japanese Speech Database", in Proceeding of EUROSPEECH, Edinburgh September 1987, vol 2, pp 13-16.
[Wagner 81] W. Wagner, "Automatic Labelling of Continuous Speech with a given Phonetic Transcription Using Dynamic Programming Algorithms", in Proceedings ICASSP-IEEE, pp 1156-9.

19es J.E.P. - Bruxelles, 19 au 22 mai 1992.

EUROM1: UNE BASE DE DONNEES "PAROLE" MULTILINGUE PARTIE FRANCAISE [1)]

J. ZEILIGER*; J-F. SERIGNAT*; D. AUTESSERRE**; J-M. DOLMAZON*.

*Institut de la Communication Parlée, Grenoble - **Institut de Phonétique, Aix-en-Provence - FRANCE

Résumé

EUROM-1 est la première base de données "Parole" européenne, car véritablement multilingue. Huit langues y sont représentées avec un matériel linguistique équivalent, enregistrées à l'aide de méthodologies et de standards communs élaborés dans le cadre du projet ESPRIT SAM n° 2589 ("Speech Assessment Methodology"). Cet article présente brièvement le projet initial, et la réalisation de la partie française de la base.

I. EUROM-1 ET SON CADRE

La standardisation et l'enregistrement de la base européenne multilingue EUROM1 est une des actions scientifiques majeures du projet SAM (ESPRIT N°2589 "Speech Assessment Methodologies") dont l'objectif est de standardiser les méthodologies et les outils pour l'évaluation des technologies vocales dans l'espace pluri-linguistique européen. Cette base doit contenir à terme et pour l'ensemble des 8 pays représentés (Italie, Angleterre, Allemagne, Pays-Bas, Danemark, Suède, Norvège, France) l'équivalent dans chacune des langues d'un ensemble de corpus, prononcés par un même nombre de locuteurs sélectionnés selon les mêmes critères, placés dans les mêmes conditions d'enregistrement.

La France, représentée par le GDR-PRC qui regroupe six laboratoires, joue dans cette action un rôle moteur et a beaucoup investi dans la définition des standards et des procédures. L'ICP, maître d'oeuvre pour le GDR-PRC de l'opération BDSONS (Base de Données des SONS du français) a participé activement à l'élaboration des standards et des outils informatiques les mettant en oeuvre (définition de l'architecture du poste de travail type standardisé au niveau européen (station SESAM)[5], logiciel d'enregistrement de corpus automatisé (EUROPEC)[2], logiciel d'édition et d'étiquetage manuel de signal de parole PTS [5], logiciel de préparation des corpus (BDPEC), logiciel de mise à jour de base de données (CHARGEBD), système de gestion de base de données GERSONS)[7]. Cette méthodologie et ces outils sont exploités à l'ICP où a été effectué l'enregistrement de la partie française d'EUROM1 pour le compte du GDR-PRC. Cette action a été menée en collaboration avec l'Institut de Phonétique d'Aix-en-Provence qui a adapté au français le contenu linguistique des corpus définis au sein de la Communauté Européenne.

II. LE CONTENU

II.1 LES CORPUS

Rappelons qu'ils constituent une adaptation française (et non pas une traduction stricte) des corpus types définis en commun [4].

I) Cette action a été menée grâce au soutien du Ministère de la Recherche et de la Technologie (MRT: départements IST et MTI), de la C.E.E. (projet ESPRIT n° 2589 "SAM" : Multi-lingual speech input/output assessment, methodology and standardisation) et du C.N.R.S.

Parole continue : - 40 passages de cinq phrases thématiquement liées, présentant une structure sémantique cohérente afin de fournir une structure prosodique correcte au niveau de chaque phrase.

- 50 phrases de complément destinées à faire apparaître les particularités du langage qui ne sont pas présentes dans les textes précédents.

Logatomes : 82 CVC (combinaison Consonne + Voyelle + Consonne) représentant les variations possibles dans la langue française de la consonne initiale et de la consonne finale avec les voyelles extrêmes [a,i,u]. Ces logatomes sont prononcés seuls ou en contexte, avec cinq contextes différents.
Les mots du contexte ont aussi été prononcés isolément.

Nombres : 5 groupes de 20 nombres (de 0 à 9999) choisis pour couvrir toutes les possibilités phonotactiques de l'ensemble de nombres considéré.

On pourra trouver une description détaillée de tous ces corpus dans le rapport final du projet SAM [4].

II.2. LES LOCUTEURS

Tous les locuteurs sont de langue maternelle française. Ils sont répartis en trois sous-ensembles: un premier groupe de 60 locuteurs ne présentant pas d'accent régional très marqué, excepté pour 5 à 6 d'entre eux; un deuxième groupe de 10 locuteurs choisis parmi les précédents et ne présentant pas d'accent régional du tout; enfin un troisième groupe de 4 locuteurs choisis parmi les 10 précédents.

Pour le deuxième groupe, le signal glotto-pharyngien a été enregistré en parallèle avec le signal de parole, au moyen d'un laryngographe.

III. CONDITIONS D'ENREGISTREMENT:

Une méthodologie standard commune à tous les sites d'enregistrement a été adoptée au sein du projet [1]. Les enregistrements de la partie francaise ont été réalisés en chambre sourde à l'ICP Grenoble.

Des "enregistrements pilotes" ont été effectués avant le début de la campagne d'enregistrement et analysés par le laboratoire NPL (Teddington, UK), afin d'évaluer la qualité de la chambre sourde et de la chaine d'acquisition. Pour l'ensemble de la chaine acoustique, le niveau de bruit de fond (silence entre les mots) est à -70 dB du niveau maximum d'enregistrement pour un niveau crête du signal à environ -10 dB.

Des procédures de test et de calibration ont été mises en oeuvre pour assurer la qualité des enregistrements à chaque étape.

Les enregistrements ont été effectués en mode dit "continu", de façon à collecter les réalisations aussi bien que les évènements extra-linguistiques (bruits de lèvres, toussotements, etc...), ceci afin de satisfaire les besoins en **parole dite naturelle.**

IV. DONNEES STATISTIQUES

La campagne d'enregistrement a duré environ 6 mois (de juin à décembre 1991), au cours desquels un peu plus d'une centaine de sessions d'enregistrement se sont tenues. Chaque session durant en moyenne deux heures, se décomposait en 1 heure d'enregistrement (locuteur dans la chambre sourde) et une heure de vérification (validité et qualité) et de sauvegardes diverses.

L'étendue des corpus a été prévue pour permettre une large couverture des particularités linguistiques et il a été décidé de ne pas faire prononcer tous les corpus par tous les locuteurs.

Dans le groupe 1, chaque locuteur a prononcé 3 passages sur les 40 par roulement, 1 groupe de 5 phrases sur les 50 par roulement et les 5 groupes de 20 nombres. Chaque passage a donc été prononcé en moyenne 4 à 5 fois, chaque phrase 6 fois, et chacun des 100 nombres 60 fois.

Dans le groupe 2, chaque locuteur a prononcé 10 passages, 5 groupes de 5 phrases, 5 fois les 100 nombres et 5 fois tous les logatomes. Chaque passage a donc été prononcé en moyenne 2 à 3 fois, chaque phrase 5 fois, chaque nombre 50 fois, et chaque logatome 50 fois également.

Dans le groupe 3, chaque locuteur a prononcé tous les logatomes dans les 5 contextes différents, et 10 fois les mots du contexte en isolé. Chaque logatome a donc été prononcé 4 fois dans chacun des contextes, et chaque mot du contexte 40 fois.

Tous locuteurs confondus, la base contient donc 280 passages, 550 phrases, 11000 nombres, 4100 logatomes, 1640 logatomes en contexte, et 400 occurrences de mots des contextes. Les tableaux suivants présentent de manière synthétique toutes ces informations.

NB: un "item" est ici un constituant de base d'un corpus, c'est donc suivant le type de corpus: 1 passage ou 1 phrase ou 1 nombre...

Groupe 1:

Groupe 1: 60 locuteurs	PASSAGES	PHRASES	NOMBRES
Chaque locuteur a prononcé:	3	5	100
Chaque item° a été prononcé en moyenne:	4 à 5 fois	6 fois	60 fois

Groupe 2:

Groupe 2: 10 locuteurs	PASSAGES	PHRASES	NOMBRES	CVC
Chaque locuteur a prononcé:	10	25	100 * 5	82 * 5
Chaque item a été prononcé en moyenne:	2 à 3 fois	5 fois	50 fois	50 fois

Groupe 3:

Gr. 3: 4 locuteurs	CVCctx1	CVCctx2	CVCctx3	CVCctx4	CVCctx5	context
chaque locuteur	82	82	82	82	82	10*10
chaque item	4 fois	4 fois	4 fois	4 fois	4 fois	40 fois

* Le volume des données

Groupe 1: 15 Mo par locuteur;
total: 15 * 60 = 900 Mo
Groupe 2: 200 Mo par locuteur;
total: 200 * 10 = 2000 Mo*
Groupe 3: 60 Mo par locuteur;
total: 60 * 4 = 240 Mo

TOTAL: 3140 Mo environ

*(dont moitié signal de parole, moitié signal de laryngographe)

* Population des locuteurs

60 locuteurs ont été retenus après sélection des 70 personnes enregistrées. Diverses données ont été recueillies concernant les locuteurs: sexe, âge, origine géographique, rapport taille/poids, niveau d'éducation, fumeur ou non... Une analyse rapide montre que l'équilibre est réalisé: 30 femmes et 30 hommes. Toutes les régions de France sont grossièrement représentées, avec deux pôles plus importants en région Rhône-Alpes et Parisienne. La répartition en tranche d'âge couvre l'intervalle de 20 à 62 ans. Trois quarts des locuteurs ont fait des études supérieures et un quart des études secondaires; aucun ne s'est arrêté au primaire. Ceci est valable pour les hommes comme pour les femmes. Tout comme l'habitude de la cigarette qui est fréquente chez 1/3 des personnes enregistrées, hommes et femmes à égalité.

CONCLUSION:

La partie française de la base de données EUROM-1 est actuellement stockée sur cassettes Exabyte. Au niveau européen, le laboratoire anglais NPL (National Physical Laboratory, Teddington, Middlesex) est chargé de l'opération de transfert de l'intégralité de la base EUROM1 sur support de type CD-ROM (standard ISO-9660). On peut espérer une disponibilité de la base dans le courant de l'année 1992.

D'autre part, le logiciel (GERSONS) développé pour la gestion de la base de données des sons du français (BDSONS) a été adapté pour permettre la gestion d'une base multlingue telle que EUROM1 [7]. Ce logiciel, doté d'une interface utilisateur conviviale, facilite l'exploitation de la base (sélection de corpus, sélection de locuteurs, extraction de sons...).

Bibliographie:

[1] M.J TOMLINSON: "Guide to Database Generation - Recording Protocol", Final version - SAM-RSRE-015 - Jan 91.
source: RSRE, Malvern, England.

[2] J. ZEILIGER - J.F. SERIGNAT: "EUROPEC software (v4.1), User's Guide Release 4.1" - SAM-ICP-045 - Mar 91.
source: Institut de la Communication Parlée, Grenoble.

[3] J.M. DOLMAZON - J.F. SERIGNAT - J.C. CAEROU - J.ZEILIGER: "Contribution to the final report (March 1989 - February 1992)" -SAM-ICP-050 - Jan 92.
source: Institut de la Communication Parlée, Grenoble.

[4] J.F. SERIGNAT - J. ZEILIGER: "Annex ICP 1 - List of EUROM1 database corpora (French Part)" - SAM-ICP-051 - Jan 92.
source: Institut de la Communication Parlée, Grenoble.

[5] J.C. CAEROU - J.M. DOLMAZON - A. EL BADMOUSSI: "Une boîte à outils "Parole" pour la station SESAM". Publié dans 19ème JEP, Bruxelles, 19992.

[6] P. FOULARD: "Gestion de bases de données de sons de parole sur poste de travail PC". Thèse ingénieur CNAM, ICP-Grenoble, 22 Nov 1991.

[7]J.F. SERIGNAT - P. FOULARD: "GERSONS: système multibase et multilingue de gestions de sons de parole", publié dans 19ème JEP, Bruxelles, Mai 1992.

19es J.E.P. - Bruxelles, 19 au 22 mai 1992.

UNE BOITE À OUTILS "PAROLE" POUR LA STATION SESAM[1]

J. C. CAEROU, J. M. DOLMAZON, A. EL BADMOUSSI

Institut de la Communication Parlée, Unité de Recherche Associée au CNRS 368, INPG et Université Stendhal, 46, avenue Félix Viallet **38031 GRENOBLE CEDEX** France.

Résumé

Les échanges sont favorisés, en Europe, par les projets mis en place depuis quelques années par la CEE. Ainsi le projet ESPRIT II 2589, Multilingual Speech Input-/Output Assessment, Methodology and Standardisation (SAM) a rassemblé les laboratoires de parole de huit pays Européens, dont les laboratoires Français groupés dans le GDR-PRC "Communication Homme Machine". Le but de ce projet était de standardiser les outils et les procédures d'évaluation des Entrées-Sorties parole. Dans ce cadre, de nombreux travaux ont été entrepris pour normaliser les méthodologies d'évaluation et les outils (matériels et logiciels) associés, les bases de données pour les corpus, les procédures d'enregistrement et d'étiquetage, les logiciels de bases, etc.).
Dans cet article, nous présentons les résultats des travaux effectués, en vue du développement d'une station de travail dédiée à l'étude de la parole, baptisée SESAM. Nous verrons l'environnement matériel, le logiciel et les normalisations. Nous présenterons plus en détail le progiciel PTS, constituant "l'outillage de base" du traitement de signal de parole sur SESAM.

I. LA PLATE-FORME

Compte tenu des objectifs du contrat SAM, l'une des premières préoccupation des contractants en 1988, a été de définir une plate-forme standard pour recevoir les logiciels et les matériels nécessaires à l'accomplissement du projet. Les principales lignes du cahier des charges de cette plate-forme étaient [VII] :

- une grande diffusion dans tous les pays concernés,
- un prix très abordable, pour être accessible à des laboratoires peu fortunés,
- d'un emploi aisé (pas de nécessité absolue d'un personnel spécialisé en informatique).
- souple sur le plan matériel. L'adjonction de cartes doit se résumer à une opération simple et rapide.

C'est ainsi que la solution des micro-ordinateurs de type PC-AT-3 ou compatibles s'est imposée à l'époque. Le type exact de machine est laissé à l'appréciation de l'acquéreur, la seule contrainte étant d'être compatible avec un IBM PC-AT, malgré la quasi-obsolescence de cette dernière machine en 1988. Le minimum requis pour la configuration de base était un disque dur, un lecteur de disquettes de 1,2 Mo, un port parallèle, deux ports série dont un avec une souris et un adaptateur graphique EGA. La carte mère devait être en mesure de recevoir un interface CD-ROM, une carte réseau Ethernet, un interface clavier "intelligent" et une carte de conversions analogique-numérique. Pour cette dernière la carte OROS AU21 s'est rapidement imposée comme le standard (avec, pour les applications bi-voies utilisation de la carte AU22).

L'évolution rapide des performances des micro-ordinateurs, ainsi que des modèles, a permis de proposer des matériels bien plus performants. La plate-forme de 1992 est basée sur un microprocesseur 80386 ou 80486, l'interface graphique est VGA (640x480) ou mieux, le minimum de RAM est de 2 Moctets, et le disque dur a une capacité d'au moins 100 Moctets.

Il reste a toutefois un point sur lequel il a été difficile de proposer des recommandations précises dès le début du projet, c'est celui d'une mémoire de masse. Les évolutions rapides de leurs performances, ainsi que les prix élevés ont paralysés le choix jusqu'en 1991. Une

[1]Définie dans le cadre du projet ESPRIT N° 2589 - SAM

étude des différentes catégories de mémoires de masses a été faite [V]. Cette étude fait ressortir que le DAT est maintenant le système de stockage le plus performant pour les très gros volume (supérieurs à 1 Goctets). Malgré ces conclusions, et à cause d'une grande diffusion dans les laboratoires, le choix pour la fin du projet SAM s'est porté sur le système Exabyte.

II. LES LOGICIELS

Dans chacune des trois tâches présentées en introduction, de nombreux outils logiciels ont été développés. La collection d'enregistrements, en vue de la création de bases de données, est régie par des protocoles standardisés, et un logiciel baptisé EUROPEC a été développé pour mettre en place ces procédures de façon entièrement automatique. Il assure sous contrôle d'un manager le suivi depuis la phase d'enregistrement, au cours de laquelle il prend en charge le locuteur, jusqu'à la constitution finale des fichiers de signaux et d'informations associées. Cet outil a été développé avec l'idée sous-jacente de la constitution aisée de bases de données de paroles pratiques et bien documentées. Ces bases sont finalement distribuées sur CD-ROM qui est un support de masse volumineux, pratique, de faible encombrement et bien adapté à la diffusion de grandes quantités (multimédia).

Pour une utilisation optimale, les bases de données doivent être étiquetées. Trois logiciels d'étiquetage semi-automatique ont été développés et testés sur plusieurs langues, chez la plupart des partenaires. Ces outils sont d'un grand intérêt pratique pour le projet en terme de préparation des bases de données et de la connaissance de base de l'étiquetage multilingue. L'évaluation de ces étiquetages est effectué par un outil d'expertise : ELSA. L'étiquetage utilise des alphabets normalisés pour la transcription ASCII des symboles phonétiques IPA : SAMPA pour l'étiquetage phonétique, SAMPROSA et SAMSINT pour les étiquetages prosodiques.

L'évaluation des systèmes de reconnaissance se fait sous contrôle d'un logiciel (EURPAC), qui assure une cohérence des tests et des résultats. Des outils complémentaires fournissent des moyens d'évaluation statistiques des résultats (SANOVA) et également l'accès et l'enrichissement des bases de données (RISE).

De manière analogue l'évaluation de la sortie parole est contrôlée par un outil qui génère des stimulus (SUSGEN), prend en charge les spécificités des synthétiseurs et analyse les résultats des auditeurs.

En plus de ces logiciels spécifiques, d'autres outils plus généraux ont été développés. C'est ainsi que des outils logiciels pour l'analyse de signaux de parole ("Software Tools for Speech Signal Analysis") ont vu le jour. Le paragraphe IV présente ces outils plus en détails.

D'autres logiciels plus spécifiques ont aussi été développés pour les différentes tâches d'évaluation. On en trouvera une description exhaustive dans le rapport final du projet [VI].

III. LES NORMES

Nous avons vu ci-dessus les nécessités de normalisation dans la définition de la station de travail SAM. Cette normalisation reste sans effet si les échanges de logiciels et de données ne peuvent pas se faire correctement. Une des conditions pour obtenir une bonne circulation des données est de normaliser les formes. Malheureusement, très souvent la notion de normalisation est opposée à la notion de souplesse et de liberté. Au sein de SAM nous avons essayé de concilier la normalisation et la souplesse.
Le premier point est la méthode d'échanges de données par fichiers. Les logiciels dont nous avons parlé ci-dessus puisent leurs entrées dans des fichiers de données ou de paramètres, tandis que les sorties sont également stockées dans des fichiers qui servent à leur tours d'entrées pour d'autres logiciels. Cette méthode, peut-être un peu lourde, offre de nombreux avantages :

- souplesse de stockage,
- facilité d'échanges,
- possibilités de conversions si nécessaire.

Le second point concerne le format des fichiers. Ils doivent être facile à exploiter, donc directement lisibles par un être humain. C'est pourquoi ils sont, pour la plupart, en format **texte ASCII** (bien entendu il y des exceptions : les échantillons des signaux de parole sont codés en entiers 16 bits signés). la nomenclature des fichiers a un rôle considérable à jouer dans leur facilité d'utilisation et de manipulation. Ce point est particulièrement crucial pour les centaines ou milliers d'enregistrements de locuteurs dans les bases de données. Une nomenclature bien pensée peut grandement faciliter certaines opérations (exemples : des requêtes). La souplesse résulte de la possibilité d'ajouter des informations non prévues lors de la définition des fichiers, grâce à l'utilisation de mots clef précédant les données. Les fichiers descripteurs associés aux enregistrements, ainsi que les fichiers d'étiquetage suivent cette syntaxe [VIII]. Ces fichiers comportent un ou plusieurs blocs eux-mêmes formés d'une en-tête et d'un corps. Les en-têtes contiennent des informations

générales (date, expert, version, laboratoires, etc.), tandis que les corps intègrent les données proprement dites (par exemple : nature de l'étiquette identifiée par un mot clef, suivie de ses coordonnées et du code ASCII). La Figure 1 montre un extrait d'un fichier d'étiquetage contenant deux blocs, avec leur en-tête et leur corps.

```
LHD:V1.1
FIL:LABEL
TYP:PHONEMIC
VOL:EUROM0
DIR:\ENGLISH\PA
SRC:PA001SEAE
TXF:
SAM:16020
BEG:413474
END:444106
EXP:LHUNT
SYS:
DAT:27/04/89
SPA:1.1
LBD:
LBB: 413474, 413554, 413634, b
LBB: 413634, 413937, 414240, @
CMT: here user can add comments
LBB: ....
.........
LBB: 440969, 442537, 444106, S
ELF:
LHD:V1.1
FIL:LABEL
DAT:27/04/89
SPA:1.1
LBD:
LBB: 450038, 450302, 450567, w
LBB: 450567, 450987, 451408, e
.....
LBB: 461127, 463911, 466696, O:
ELF:
```

Figure 1 : Un extrait de fichier d'étiquetage.

La normalisation concerne également la manière d'écrire les logiciels. La restriction du nombre de langage favorise leur connaissance approfondie, et de meilleures possibilités d'échange. Dans le projet SAM, le langage principalement utilisé est le C Microsoft version 5.1. L'assembleur a parfois été utilisé dans des sections critiques. Les programmes sur la carte OROS sont programmés en assembleur TMS320C25 et en OPAL (un langage structuré générant de l'assembleur TMS320C25).

Pour assurer une uniformité dans le développement des logiciels, les règles de développement ont été éditées et distribuées aux partenaires du projet [III]. Ce document rappelle non seulement les règles de production des logiciels avec la normalisation du code C, mais aussi la nécessité de penser le logiciel en termes de testabilité, de documentation et de maintenance.

IV. UNE BOITE À OUTILS POUR L'ANALYSE DU SIGNAL DE PAROLE : PTS

Dans cette dernière partie nous mettons plus particulièrement en évidence les caractéristiques et les originalités du progiciel de traitement de parole "PTS". Il constitue le "basic soft package" de la station SESAM. Le but de PTS est de fournir aux utilisateurs un jeu d'outils de visualisation d'écoute et de manipulation du signal de parole, ainsi que des résultats de traitements faits sur ces même signaux. PTS travaille dans un environnement graphique, utilisant des fenêtres pour afficher les différent objets. On peut par exemple visualiser la forme d'un signal de parole au cours du temps. Des curseurs verticaux permettent de délimiter une portion qui peut être "zoomée", écoutée, traitée pour visualiser le sonagramme correspondant dans une autre fenêtre, etc.

Deux types d'objets sont manipulés par PTS :

- **Les signaux.** Ce sont les résultats des enregistrements. PTS utilise bien entendu les formats de fichiers de parole définis dans le projet SAM (un fichier de signal accompagné d'un fichier descripteur). Mais il sait également reconnaître et lire d'autres formats de fichiers de signal (ILS...).

- **Les paramètres.** Ces derniers contiennent tous les types de données à l'exception de signaux. Ainsi on peut imaginer de stocker dans un même fichier de paramètres les données résultant du calcul d'un sonagramme, la courbe caractéristique de F0, la courbe d'énergie, et éventuellement les formants calculés sur les voyelles. Des en-têtes fournissent les informations nécessaires à l'extraction de l'un ou de l'autre de ces paramètres. Ces possibilités sont particulièrement intéressantes dans le cadre des réseaux : tout traitement effectué sur une autre machine est directement utilisable dans PTS par le simple transfert d'un fichier de paramètre.

PTS est doté de nombreuses fonctions de visualisation, d'écoute et d'édition des signaux de parole. Dans le cas des fichiers de sons au format SAM, l'utilisateur peut visualiser l'ensemble d'un fichier ou ne sélectionner qu'un des enregistrement. Le contenu de la fenêtre de visualisation peut subir des "zoom" verticaux et horizontaux divers afin d'adapter l'objet à visualiser aux besoins de l'utilisateur. Pour les écoutes, les concepts sont du même type : écoutes totales ou sélectives. Ces

dernières fonctions, ainsi que quelques autres calculs (sonagrammes) nécessitent que la station soit dotée de la carte OROS, qui joue deux rôles : ses convertisseurs assurent les fonctions d'entrées/sorties des signaux analogiques, tandis que son DSP s'occupe de fonctions de traitement de signal (FFT dans le cas du sonagramme).

PTS est doté des traitements classiques du signal de parole : la représentation des sonagrammes, la coupe du sonagramme à un instant donné, le calcul de F0 par AMDF [I], le calcul de l'énergie. Comme pour les autres fonctions, PTS propose un paramétrage des fonctions de calculs. Ainsi, par exemple chaque utilisateur peut se constituer son propre fichier de configuration des calculs et de la représentation des sonagrammes (large bande, bande étroite, dynamique). Plutôt que de multiplier les fonctions de traitement au sein de PTS, ce qui le rendrait gigantesque, nous avons préféré privilégier l'ouverture sur le monde extérieur par le concept des fichiers de paramètres présentés ci-dessus. Un complément indispensable à ce dernier concept est l'accès aux fonctions de réseaux qui autorisent l'échange de fichiers et la soumission de tâches sur d'autres machines. On a ainsi un accès aisé à des traitements non implémentés dans PTS.

Les facilités de visualisation du type duplications de fenêtres, synchronisations de curseurs, choix de palette de couleurs, etc. sont appliquées à tous les types de contextes.

PTS est un puissant outil d'étiquetage manuel. Une interface utilisateur conviviale et un environnement très ouvert, permettent tout type d'étiquetage temporel au niveau du signal (marques phonétiques, prosodiques, linguistiques, etc.). La visualisation simultanée sur un contexte (signal, sonagramme ou autre) des étiquettes en cours et d'un étiquetage de référence, offrent une grande souplesse à un utilisateur pour expertiser les résultats d'un autre étiquetage et d'y apporter les corrections éventuelles. Cette possibilité est indispensable pour valider les résultats des étiquetages automatiques ou semi-automatiques (c.f. paragraphe II).

Comme pour les autres fonctions de PTS nous avons tenu à conserver le maximum de souplesse en offrant aux utilisateurs des fichiers de configuration : la définition des types d'étiquettes, ainsi que les relations entre symboles graphiques et code(s) ASCII correspondant(s), sont définis dans des fichiers configurables par l'utilisateur. Pour ceux qui préfèrent les étiquettes avec une représentation graphique, un éditeur permet de créer et modifier des jeux de symboles utilisables dans l'environnement de PTS.

Des mesures sont offertes sur tous les types de visualisations. Par l'utilisation de curseurs ou de croix, la manière d'opérer est intuitive. Les résultats peuvent être stockés dans des fichiers pour une exploitation ultérieure.

L'ensemble des fonctions dont nous avons esquissé ci-dessus un petit tableau, occupe un volume important dans la mémoire de la station (il faut garder présent à l'esprit que sous MS-DOS nous ne disposons que de 640 Koctets de mémoire conventionnelle). Pour nous libérer des limites de la mémoire conventionnelle, nous avons expérimenté et décidé d'utiliser la mémoire étendue. Malheureusement les arcanes de l'espace mémoire des PC sont complexes (UMB, EMS, XMS, extended memory, etc.), et les modes de fonctionnement du processeur diffèrent selon l'endroit où l'on se situe. Après une brève expérimentation de la mémoire paginée (EMS) en mode réel pour tenter d'y loger les gros fichiers d'étiquettes, nous avons décidé un portage de PTS en mode protégé dans l'espace mémoire étendu (au delà de 1 Moctets). Ce portage a été rendu possible grâce à un "DOS Extender". Outre le gain de place (on dispose maintenant d'une taille pouvant aller jusqu'à 16 Moctets, en fonction des ressources de la station), le fonctionnement en mode protégé offre des protections logicielles, soit un confort accru pour l'utilisateur.

Pour terminer avec ce chapitre, signalons que PTS est muni d'une notice d'utilisation permettant aux utilisateurs d'exploiter pleinement les fonctions [IV]. Elle comporte dans les annexes, les descriptions techniques des fichiers de configuration dont nous avons fait mention tout au long de la description du logiciel. L'utilisateur peut ainsi configurer PTS à sa manière.

CONCLUSION

Nous avons décrit dans cet article, les éléments constitutifs d'une station de travail pour la parole, définie et utilisée dans le cadre d'un projet Européen ESPRIT. L'objectif primaire était le bas prix, la grande disponibilité, la modularité et la facilité d'emploi. Nous avons montré les solutions retenues pour le matériel, et comment l'évolution rapide du marché de la micro-informatique nous a conduit à les réviser. Les nombreux partenaires du projet ont su maintenir une ligne de conduite pour s'imposer des normes, évitant ainsi une anarchie dans les développements des logiciels, tout en conservant une souplesse suffisante pour éviter les scléroses. Dans la dernière partie de l'article nous avons présenté les principales caractéristiques du logiciel de base, qui constitue un complément indispensable à la

station SESAM. C'est la raison de sa dénomination "boite à outils pour l'analyse du signal de parole".

Le travail présenté dans cet article est le fruit d'une intense collaboration entre les partenaires du projet ESPRIT n° 2589 (SAM) "Multilingual Speech Input-/Output Assessment, Methodology and Standardisation". Les auteurs remercient tous ceux qui ont participé à la définition et l'implémentation des concepts de base de cette station de travail.

BIBLIOGRAPHIE

[I] T. BARBE, G. BAILLY, Evaluation d'un détecteur de fréquence fondamentale du signal microphonique par comparaison à une référence laryngographique, XVIIIèmes Journées d'Études de la Parole, Montréal (Québec), Canada, 28-31 mai 1990, pp. 165-169.

[II] J. C. CAEROU, J.M. DOLMAZON, J. M. LUNATI, SESAM: a Low Cost Workstation for Speech Assessment, Proceedings of ESCA Tutorial Day and Workshop on Speech Input/Output Assessment and speech Databases, Noordwijkerhout, the Netherlands, 20-23 september 1989.

[III] J.C. CAEROU & Al., Definition of Software Production, Documentation, etc., Esprit Project 2589 (SAM), Multi-lingual Speech Input/Output Assessment, Methodology and standardisation January 11th 1990.

[IV] J.C. CAEROU & Al., PTS SOFTWARE V:4.30, USER MANUAL, Esprit Project 2589 (SAM), Multi-lingual Speech Input/Output Assessment, Methodology and standardisation, June, 13th 1991.

[V] J.C. CAEROU, J.M. DOLMAZON, Mass Data Storage: Review of the present State and first draft of Proposal, esprit project 2589 (SAM), multi-lingual speech input/output assessment, methodology and standardisation, SAM-ICP-047, june 1991.

[VI] ESPRIT PROJECT 2589 (SAM), Multilingual Speech Input/Output Assessment, Methodology and Standardisation, FINAL REPORT, April 1992.

[VII] B. LINDBERG, S.W. DANIELSEN, Specification of the Low Level SESAM, Esprit project 1541 (1541), Multi-lingual Speech Input/Output Assessment, Methodology and standardisation, Extension Phase Final Report, 28 February 1989, pp. 210-221.

[VIII] M. TOMLINSON, R. WINSKI, B. BARRY, Label File Format Proposal, Esprit project 1541 (1541), Multi-lingual Speech Input/Output Assessment, Methodology and standardisation, Extension Phase Final Report, 28 February 1989, pp. 189-197.

19es J.E.P. - Bruxelles, 19 au 22 mai 1992.

ANALYSE-SYNTHESE PAR DECOMPOSITION DE LA PARTIE DETERMINISTE ET DE LA PARTIE ALEATOIRE DU SIGNAL DE PAROLE

Sophie Grau & Christophe d'Alessandro

LIMSI-CNRS, BP 133, 91403 Orsay Cédex

RÉSUMÉ

Une méthode décomposition du signal de parole en une partie quasi-harmonique et en une partie aléatoire résiduelle, s'inspirant de la thèse de Serra [7] est décrite. Cette méthode est un développement du codage sinusoïdal, en imposant des contraintes supplémentaires sur les composantes sinusoïdale afin de pouvoir les interpréter en termes du modèle acoustique de production de la parole. L'article décrit les étapes de l'algorithme d'analyse: calcul des spectres à court terme, extraction sous contrainte des pics, définition des trajectoires spectrales, et l'algorithme de synthèse, qui délivre une parole synthétique de qualité transparente. L'application de cette méthode à l'analyse de la parole est ensuite discutée. Il apparait que la séparation obtenue est facilement interprétable en terme acoustiques, et doit permettre l'étude de phénomènes difficiles à apréhender par d'autres méthodes: excitation mixte, bruit d'aspiration par exemple.

Mots clés: *représentation sinusoïdale, représentation harmonique, modification de parole*

1 Introduction

Les méthodes d'analyse-synthèse fondées sur une représentation harmonique ou sinusoïdale suscitent beaucoup d'interêt depuis quelques années, en particulier dans le domaine du codage de la parole. Deux classes de modèles peuvent être distinguées : le modèle sinusoïdal [4] dans lequel un signal est représenté comme une somme de sinusoïdes variant dans le temps en amplitude, fréquence et phase, et le modèle harmonique [1] dans lequel des contraintes d'harmonicité sont ajoutées sur les trajectoires des sinusoïdes.

En partant du modèle sinusoïdal, Serra [7] a développé un système d'analyse/synthèse fondé sur la décomposition d'un son en une partie déterministe, qui représente la partie quasi-harmonique du son, et une partie résiduelle, qui représente la partie aléatoire. Le propos de cet article est de montrer qu'un tel système, originellement conçu pour l'analyse de sons musicaux, peut être d'une aide très précieuse pour l'étude acoustique et la synthèse de la parole.

En effet, dans le cas de la parole, ce système permet de séparer la partie voisée du signal de la partie bruitée, la partie voisée étant représentée comme une somme de composantes sinusoïdales en relations quasi-harmoniques. Nous avons ainsi testé ce système sur un corpus de parole voisée, représentatif des mélanges possibles de sources d'excitation voisée/non voisée et de la non-stationarité due aux évolutions rapides: mouvements rapides de co-articulation, bruit de souffle des fricatives voisées, bruit d'aspiration rencontré dans certaines voyelles, explosions des plosives voisées, modification importante de la fréquence fondamentale en cours d'énoncé.

Après avoir présenté le modèle acoustique sur lequel se fonde la représentation, nous détaillerons les étapes du processus d'analyse/synthèse puis nous présenterons les résultats de l'évaluation de ce système sur un corpus de parole voisée et son application à la modification de parole.

2 Modèle quasi-harmonique de la parole voisée

2.1 Relation avec la représentation sinusoïdale

Dans ce modèle le signal de parole $s(t)$ est considéré comme la somme d'une série quasi-harmonique

de sinusoïdes (partie déterministe), et d'un résiduel $e(t)$:

$$s(t) = \sum_{r=1}^{R} A_r(t) cos[\theta_r(t)] + e(t) \quad (1)$$

où R est le nombre de sinusoïdes, $A_r(t)$ est l'amplitude instantanée et $\theta_r(t)$ la phase instantanée de la rième sinusoïde.

La phase instantanée est définie de la même manière que dans un modèle sinusoïdal, c'est à dire :

$$\theta_r(t) = \int_0^t \omega_r(\tau) d\tau + \phi_r(t) \quad (2)$$

où $\omega_r(t)$ est la fréquence instantanée, qui donne les variations rapides de phase. $\phi_r(t)$ représente les variations lentes de phases dues à l'évolution du spectre.

Par rapport à un modèle sinusoïdal classique, les différences sont les suivantes :

1. les sinusoïdes sont restreintes aux composantes quasi-harmoniques de la parole voisée.
2. un signal résiduel est obtenu comme différence entre le signal original et la partie quasi-harmonique.

2.2 Relation avec le modèle acoustique source/filtre

Le modèle source/filtre de production de la parole s'écrit dans le domaine fréquentiel, pour la parole strictement voisée [2]:

$$S(\omega) = E(\omega)G(\omega)V(\omega)R(\omega) \quad (3)$$

ou E représente la quasi-périodicité de l'excitation (de fréquence $F0$), G la contribution de l'onde de débit glottique, V celle du conduit vocal, R un terme de rayonnement. Si l'on reporte 3 dans 1, on peut séparer sur le modèle quasi-harmonique les contribution des composantes acoustiques:

$$\begin{cases} A_r(t) = A_r^G(t) A_r^V(t) \\ \theta_r(t) = \int_0^t 2r\pi F0(\tau) d\tau + \phi_r^G(t) + \phi_r^V(t) \end{cases} \quad (4)$$

en associant dans G la contribution (constante) du terme de rayonnement. Cette écriture est la base de l'analyse et des modifications de la parole en termes sinusoïdaux.

3 Système d'analyse-synthèse

Serra [7] a proposé un système d'analyse/synthèse, s'appuyant sur l'utilisation de la transformée de Fourier à court-terme (STFT), et permettant d'obtenir les 2 composantes précédentes (partie déterministe et partie résiduelle) d'un son musical. Le processus d'analyse est le suivant:

1. calcul du spectre d'amplitude et de phase par la STFT.
2. extraction des pics proéminents dans chaque spectre.
3. organisation des pics en trajectoires fréquentielles grâce à un algorithme d'appariement qui sélectionne les pics souhaités.

Le signal comportant les sinusoïdes organisées en trajectoires peut ensuite être regénéré par synthèse additive, et le signal résiduel est obtenu en soustrayant spectralement le signal déterministe au signal original. Nous allons considérer les étapes d'analyse.

3.1 détection des pics spectraux

Une fois que le spectre complexe à court terme est calculé par la STFT pour chaque trame d'analyse, puis converti en coordonnées polaires, le système en extrait les pics proéminents. Un pic est défini comme un maximum local dans le spectre d'amplitude calculé pour une trame l. Cependant, tous les pics n'ont pas la même proéminence et la sélection de ces pics va dépendre de leur hauteur (en dB) par rapport aux vallées avoisinantes, vallées définies comme les plus proches minima de chaque côté du pic. De plus, puisque les pics de même hauteur n'ont pas la même pertinence perceptive, il est nécessaire de spécifier à l'algorithme de détection de pics les plages d'amplitude et de fréquence dans lesquelles la recherche des pics aura lieu. Une fois les pics sélectionnés, une interpolation parabolique est réalisée en utilisant les trois échantillons spectraux encadrant l'échantillon d'amplitude maximale, afin d'estimer de façon plus précise la véritable position du pic (et donc son amplitude et sa phase).

3.2 Appariement temporel des pics spectraux

Une fois que les pics spectraux ont été détectés, un sous-ensemble de ceux-ci sont organisés en trajectoires de pics par l'algorithme d'appariement des pics, chaque trajectoire représentant idéalement un harmonique. Toutes les trajectoires obtenues pourront ensuite être resynthétisées pour obtenir la partie déterministe du signal original. Cet algorithme nécessite que l'utilisateur ait quelques connaissances a priori sur le son à analyser, afin de régler au mieux les paramètres d'analyse. L'idée

de base de cet algorithme est l'utilisation d'un ensemble de guides fréquentiels qui progressent dans le temps à travers les pics spectraux sélectionnés pour chaque trame, en cherchant des pics appropriés (selon des contraintes spécifiées) pour former des trajectoires. L'état instantané de ces guides, c'est à dire leur fréquence, est gardé dans les variables f_1, f_2, ..., f_p, où p désigne le nombre de guides existant à cet instant-là. Les valeurs de ces variables sont continuellement mises à jour au fur et à mesure que les guides "naissent", "progressent" et "meurent". L'algorithme initial qui est assez complexe puisqu'il a été conçu pour travailler sur une grande variété de sons, peut être simplifié dans le cas de sons harmoniques, et en particulier dans le cas de la parole. En effet les guides fréquentiels sont réduits à chercher les fréquences harmoniques de la fréquence fondamentale de la trame courante, chaque guide recherchant un numéro d'harmonique précis, et le nombre de guides reste constant pendant toute la durée du son (on ne "démarre" pas de nouveaux guides, et les guides ne "meurent" pas, contrairement à l'algorithme général). La description de l'algorithme que nous faisons ici correspondra donc au cas de traitement des sons harmoniques.

Grâce à des paramètres de contrôle, modifiables par l'utilisateur, l'algorithme peut fournir des contraintes différentes sur le signal traité. Ces paramètres concernent principalement : la fréquence fondamentale initiale (approchée) du signal, la plage fréquentielle dans laquelle cette fréquence fondamentale doit être cherchée, la distance fréquentielle maximale permise entre un guide et le pic sélectionné, le nombre de guides utilisés (c'est à dire le nombre d'harmoniques recherchés).

Pour chaque trame d'analyse, les étapes de l'algorithme d'appariement sont les suivantes :

1. détection de la fréquence fondamentale;
2. mise à jour des valeurs fréquentielles des guides;
3. avancement des guides en appariant de nouveaux pics aux trajectoires correspondantes.

Avant d'avancer les guides à travers la trame n, un simple algorithme recherche la fréquence fondamentale de la trame n. Si cette fréquence $F0$ est trouvée, les valeurs des guides sont repositionnées à la série harmonique de ce nouveau $F0$. Si aucune fréquence fondamentale n'est trouvée, les guides gardent les valeurs fréquentielles résultant de la trame précédente. L'avancement des guides se fait de la manière suivante : chaque guide progresse à travers la trame n en cherchant le pic qui est le plus proche en fréquence de sa valeur courante. Le $r$ième guide (de valeur fréquentielle f_r) réclame donc la fréquence g_i telle que $|f_r - g_i|$ soit minimum. Cette différence fréquentielle doit être inférieure à la distance maximale permise afin que l'appariement soit possible entre le pic trouvé et la $r$ième trajectoire (ou harmonique) correspondant au $r$ième guide. Les situations possibles sont les suivantes :

1. Si un appariement est trouvé, le guide peut "progresser" (à moins qu'il y ait un conflit à résoudre). Le pic sélectionné est incorporé à la trajectoire correspondant au guide.

2. Si aucun appariement n'est trouvé, la trajectoire correspondant au guide doit "s'arrêter" en entrant dans la trame n. Cette trajectoire gardera la même fréquence mais son amplitude diminuera progressivement jusqu'à 0 pendant la durée d'une trame de synthèse. Cependant la valeur du guide n'est pas affectée, et la trajectoire pourra "redémarrer" ultèrieurement.

3. Si un guide trouve un appariement déjà réclamé par un autre guide, le pic est attribué au guide qui est le plus proche en fréquence, et le "perdant" recherche un autre appariement. Si le guide courant perd le conflit, il recherche simplement le meilleur pic disponible non-conflictuel. Si le guide courant gagne le conflit, il appelle récursivement l'algorithme d'appariement pour le compte du guide "délogé". Lorsque le guide "délogé" retrouvera ce pic conflictuel, il verra qu'il est perdant et cherchera un autre pic.

Ce processus est répété pour chaque guide, en résolvant les conflits récursivement, jusqu'à ce que tous les appariements possibles aient été faits.

3.3 synthèse déterministe

L'algorithme d'appariement des pics délivre les valeurs des pics correspondant aux composantes quasi-harmoniques, organisés en trajectoires. Chaque pic est représenté par un triplet $(A_r^l, \omega_r^l, \varphi_r^l)$ où l est le numéro de la trame et r le numéro de trajet (harmonique) auquel il appartient. Le procédé de synthèse calcule une trame l de composantes déterministes $d^l(n)$ par :

$$d^l(n) = \sum_{r=1}^{R^l} A_r^l cos[m\omega_r^l + \varphi_r^l] \qquad (5)$$

avec $m = 0,1,2,\ldots,S-1$

où R^l est le nombre de trajectoires présentes à la trame l et S la taille de la trame de synthèse.

Cependant, si l'on veut éviter des problèmes de raccordement aux frontières des trames, il faut interpoler les paramètres $(A_r^l, \omega_r^l, \varphi_r^l)$ d'une trame l à

la suivante. Pour le calcul de l'amplitude instantanée, une simple interpolation linéaire sera faite entre les valeurs d'amplitudes A_r^l et A_r^{l+1} de 2 trames successives l et $l+1$. Par contre, on ne peut pas interpoler simplement et indépendemment la fréquence et la phase d'une composante, car elles contrôlent toutes les deux la phase instantanée $\theta(m)$ définie comme : $\theta(m) = m\omega + \varphi$. On utilisera alors comme fonction d'interpolation une fonction cubique polynomiale afin de calculer la phase instantanée en chaque point m [4].

Finalement, l'équation de synthèse de la partie déterministe pour la trame l devient :

$$d^l(n) = \sum_{r=1}^{R^l} A_r^l(m) cos[\theta_r^l(m)] \qquad (6)$$

Le signal déterministe final résulte simplement de la juxtaposition de toutes ces trames de synthèse.

3.4 Calcul de la partie résiduelle

Le signal obtenu par la synthèse déterministe reproduit la phase instantanée et l'amplitude des harmoniques du son original. Il est donc possible de soustraire spectralement ce signal au signal original pour obtenir la partie résiduelle.

4 Applications

4.1 Analyse de la parole voisée

Le codage de parole par la somme des partie quasi-harmonique et résiduelle donne évidemment une qualité transparente. Dans le but d'évaluer l'intérêt d'une telle méthode de décomposition pour l'analyse de la parole, un corpus de parole voisée a été constitué. Il est formé de phonèmes isolés (voyelles orales extrêmes avec une fréquence fondamentale évoluant normalement, voyelles orales extrêmes avec une fréquence fondamentale évoluant rapidement, voyelles nasales), de logatomes VV (/ai/, /ia/, /au/ ..) et de logatomes VCV formés par une consonne voisée (ou une semi-voyelle) en contexte vocalique /a/, /i/, /u/. Ce corpus a été enregistré par un locuteur féminin et par des locuteurs masculins.

Les remarques qui ressortent de ces analyses du point de vue perceptif et/ou spectrographique sont les suivantes :

1. La partie quasi-harmonique resynthétisée est très proche perceptivement du signal original, en particulier pour les voix (masculine ou féminine) ne comportant que peu de bruit d'aspiration. Le timbre et l'intonation sont très bien préservés.

2. Pour une voix qui comporte beaucoup de bruit d'aspiration (les voix féminines sont en généralement plus bruitées que les voix masculines [3]), la partie harmonique sonne de façon plus "métallique" que le signal original, le bruit d'aspiration faisant partie du caractère naturel de ce type de voix.

3. Le bruit d'aspiration est très bien séparé du reste du signal par le système d'analyse/synthèse, et se retrouve dans le signal résiduel. On peut ainsi l'analyser directement.

4. Lorsque les mouvements formantiques sont rapides (cas des semi-voyelles), les formants sont marqués par leur contribution sur la phase des harmoniques. Le saut de phase à la résonance peut même apparaître dominant par rapport aux mouvements de phase dus à $F0$. De plus les mouvements formantiques seront plus visibles dans la partie résiduelle.

5. Les variations spectro-temporelles rapides (frontières nettes des occlusives nasales, explosion des plosives) tendent à être lissées dans la partie déterministe, à cause de l'interpolation des paramètres entre deux trames successives d'analyse. Ces éléments se retrouvent alors presque intégralement sur le signal résiduel.

6. La partie résiduelle possède en général une intonation différente de celle de la partie déterministe. Ceci s'explique par le fait que dans la partie déterministe l'intonation provient de la structure harmonique du signal (c'est à dire de la fréquence fondamentale), alors que dans la partie résiduelle l'intonation perçue provient des mouvements formantiques lorsque ceux-ci varient de façon importante : c'est le cas spécialement pour les semi-voyelles (logatomes /iui/, /uyu/) et pour les logatomes VV dont l'intonation ne varie pas dans le même sens que les mouvements formantiques. Par exemple dans un segment /ia/ avec une fréquence fondamentale montante, la partie quasi-harmonique possède une intonation plus montante que le signal complet, et la partie résiduelle une intonation descendante qui est liée au mouvement descendant de $F2$ (voir figures 1 et 2).

7. le bruit fricatif des fricatives voisées peut être séparé du voisement. D'une façon générale, les segments d'excitation mixte donnent des contributions quasi-harmoniques et résiduelles qui contiennent chacune un des aspects de l'excitation.

4.2 Modification du signal

La description en termes quasi-harmoniques du modèle source/filtre de production permet de modifier aisément le signal. Les modifications de $F0$ et de durées sont rapportées dans [5] [7]. On peut de plus modifier directement les contributions associées au conduit vocal et à la source de voisement.

5 Perspectives

Le système d'analyse/synthèse décrit dans cet article permet une décomposition du signal en une partie quasi-harmonique et en une partie résiduelle dont la qualité de codage est transparente. Une telle décomposition offre un outil original pour l'étude acoustique de la parole, car elle permet de distinguer les différentes sources d'excitation acoustiques qui se mélangent lors de la production. Il est clair que l'étude détaillée de ces sources est un enjeu actuellement important, par exemple pour la synthèse de parole. Le développement de cette étude vise à définir un synthétiseur à formant à partir de la représentation quasi-harmonique et de la synthèse du bruit par ondelettes aléatoires [6]. Pour la synthèse par unités concaténées, une étude est également en cours pour tirer avantage du raccordement très précis des unités grâce aux paramètres sinusoïdaux.

Références

[1] L. B. ALMEIDA, J. M. TRIBOLET, 1983. "Nonstationary modeling of voiced speech", IEEE transactions on ASSP, Vol. ASSP-31, No. 3, Juin 1983, pp. 644-678.

[2] G. FANT, 1970. **Acoustic theory of speech production** Mouton, Den Hague-Paris.

[3] KLATT, D. H., & KLATT, L. C. (1990). "Analysis, synthesis and perception of voice quality variations among female and male talkers". JASA, Vol. 87, No. 2., Feb. 90, pp. 820-857.

[4] R. J. MCAULAY, T. F. QUATIERI, 1986. "Speech analysis/synthesis based on a sinusoidal representation", IEEE transaction on ASSP, Vol. ASSP-34, No. 4, Août 1986, pp. 744-754.

[5] T. F. QUATIERI, R. J. MCAULAY, 1986. "Speech transformations based on a sinusoidal representation", IEEE transaction on ASSP, Vol. ASSP-34, No. 6, Décembre 1986, pp. 1449-1464.

[6] G. RICHARD, C. d'ALESSANDRO & S. GRAU, 1992. "Synthèse de bruit par fonctions d'onde formantiques aléatoires", dans ces mêmes actes.

[7] X. SERRA, 1989. **A system for sound analysis/ transformation / synthesis based on a deterministic plus stochastic decomposition.** thèse de PhD, CCRMA, Université de Stanford.

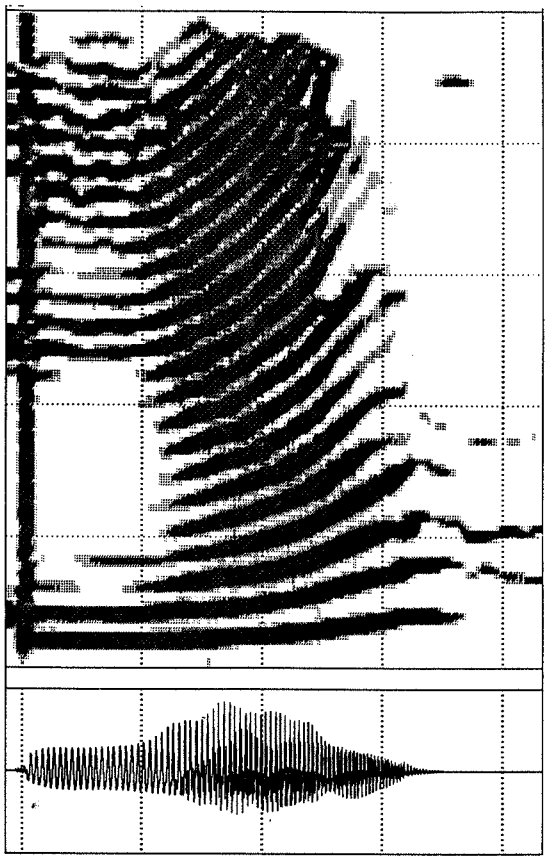


Figure 1: Segment [ia] avec une intonation montante, voix féminine: partie quasi-harmonique.

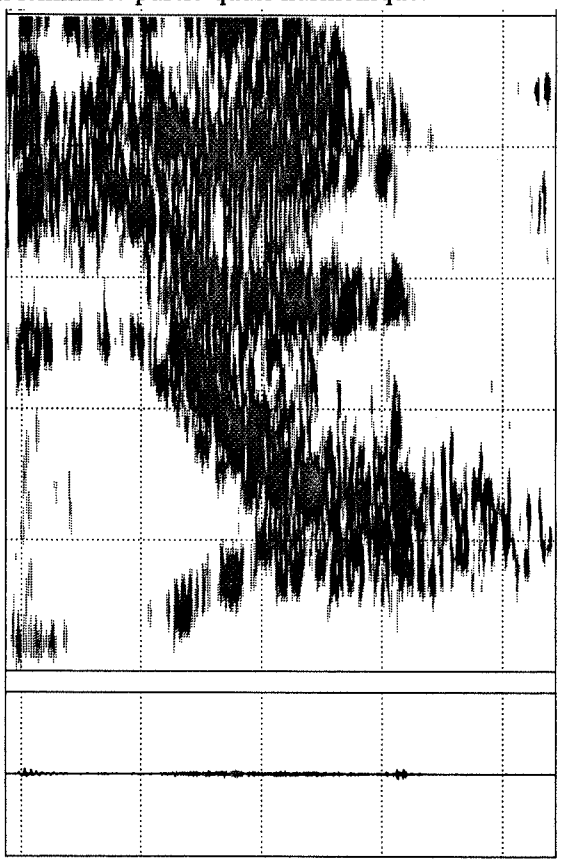


Figure 2: Segment [ia] avec une intonation montante, voix féminine: partie aléatoire.

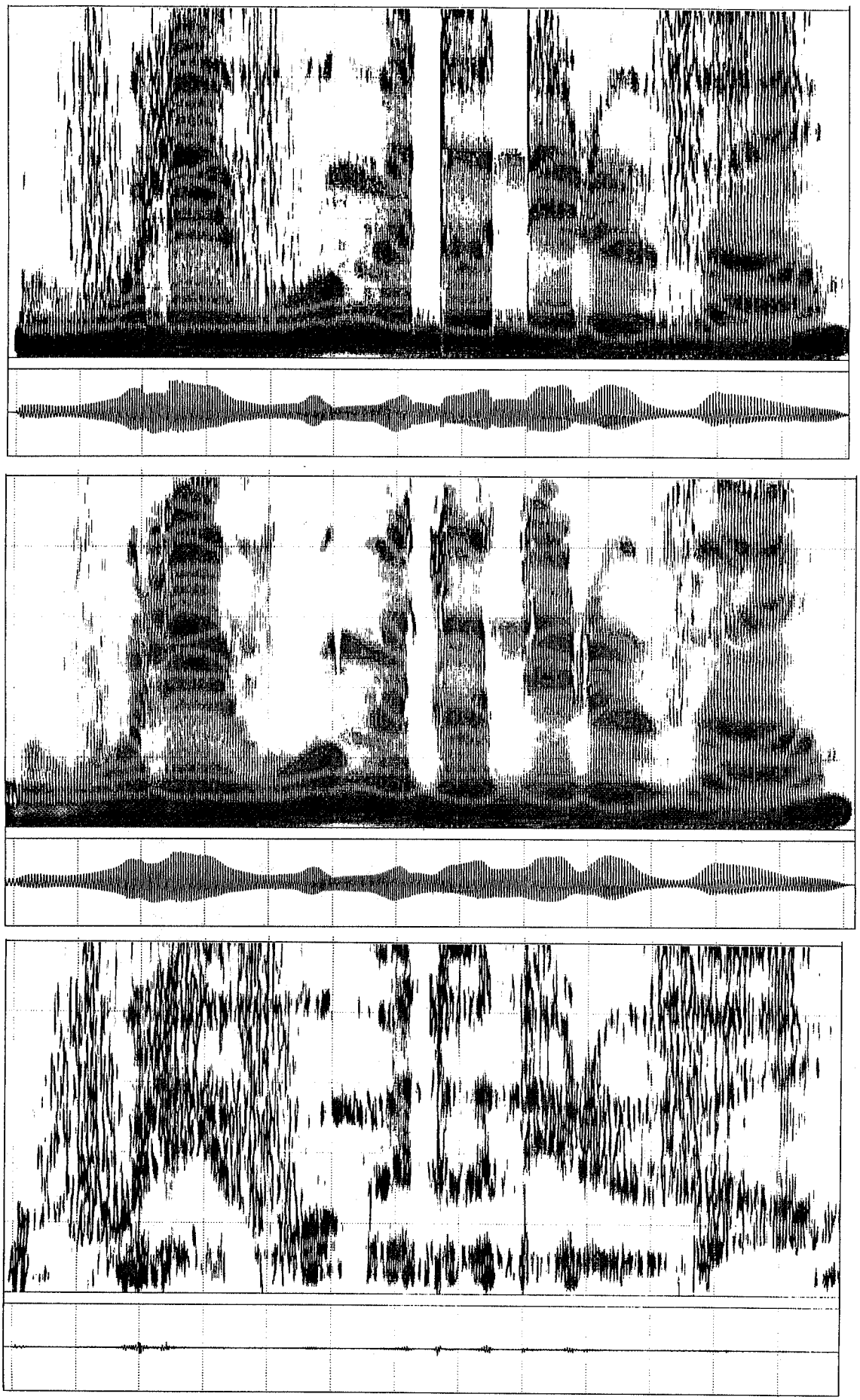


Figure 3: "Voulez-vous me donner le vin?", voix féminine. (A) original, (B) partie quasi-harmonique, (C) partie résiduelle.

19es J.E.P. - Bruxelles, 19 au 22 mai 1992.

SYNTHÈSE À PARTIR DU TEXTE D'UN VISAGE PARLANT FRANÇAIS

P. WOODWARD, T. MOHAMADI, C. BENOIT & G. BAILLY

INSTITUT DE LA COMMUNICATION PARLÉE, U.A. CNRS N° 368,
INPG/ENSERG - Université STENDHAL, BP 25X - 38040 GRENOBLE, FRANCE

Résumé

Une analyse antérieure a montré qu'une vingtaine de visages prototypiques (ou visèmes) permet de rendre compte de la diversité des formes maxillo-labiales de la parole non expressive en français. Cet article décrit les choix scientifiques et techniques pour lesquels nous avons opté : l'animation de visage par affichage d'images successives ; le jeu d'images-clé retenu ; l'architecture du système de synthèse audio-visuelle de la parole à partir du texte ; l'implémentation des régles de transcription phonème-visème tenant compte de la coarticulation en français ; et la synchronisation du son et des images synthétiques.

Ce système fonctionne sur un PC muni d'une carte de traitement du signal audio et d'une carte graphique VGA. Les règles décrites ont été élaborées sous le formalisme du compilateur de règles COMPOST, développé auparavant pour la synthèse acoustique à partir du texte.

1. INTRODUCTION

Parmi les théories et les techniques de la perception, la vision s'est affirmée d'emblée comme un domaine de recherche privilégié, et ce pour plusieurs raisons. De tous nos sens, le sens de la vue est celui qui offre la plus grande flexibilité et nous apporte la plus grande quantité d'information sur ce qui nous entoure du fait de sa triple dimension spatiale, énergétique et temporelle. La perception de notre espace, la détection du mouvement des objets qui le composent sont des données humaines essentielles et sont toutes transmises par l'intermédiaire de nos yeux.

Dans le domaine de la communication, l'omniprésence de l'informatique dans notre société fait que nous sommes toujours à la recherche d'une meilleure communication homme-machine. C'est en partie dans cette optique que l'on a commencé à équiper les ordinateurs de voix de synthèse et que l'on manifeste aujourd'hui un intérêt croissant à l'égard de la synthèse de visage parlant. Cette synthèse consiste à doter l'ordinateur d'une animation d'images faciales qui accompagne la voix artificielle.

L'animation de visage est d'autant plus nécessaire que l'intelligibilité actuelle des synthétiseurs de parole est encore très éloignée de celle des humains, et qu'il est bien connu que la perception visuelle de la parole dégradée vient compléter sa perception acoustique par un apport d'information supplémentaire (Sumby et Pollak, 1954 ; Neely, 1956 ; Binnie et al., 1974 ; Erber, 1969, 1975 ; Summerfield, 1979 ; Mohamadi et Benoît, ce volume).

En outre, il est séduisant de viser à améliorer l'impact d'un message en l'*humanisant* au moyen d'un visage, même schématique.

Dans un premier temps, nous avons élaboré une animation de visage basée sur le principe classique de l'affichage d'images successives, celles-ci étant préstockées dans un *dictionnaire de visages*. Cette technique facilite la validation des règles de transcription, bien qu'elle limite les possibilités de visualisation à un jeu fini d'expressions. Notre synthétiseur ne peut donc rendre compte, dans son état actuel, que d'une prononciation *neutre*, inexpressive, qui est néanmoins la condition optimale pour mener des tests de perception de la parole, *et seulement* de la parole...

Le synthétiseur de visage décrit ici vise donc à apporter intelligibilité et agrément à la synthèse de la parole à partir du texte, et à offrir un outil de validation des connaissances en matière de production et de perception de la parole. L'évaluation de son intelligibilité fera l'objet d'une prochaine étude.

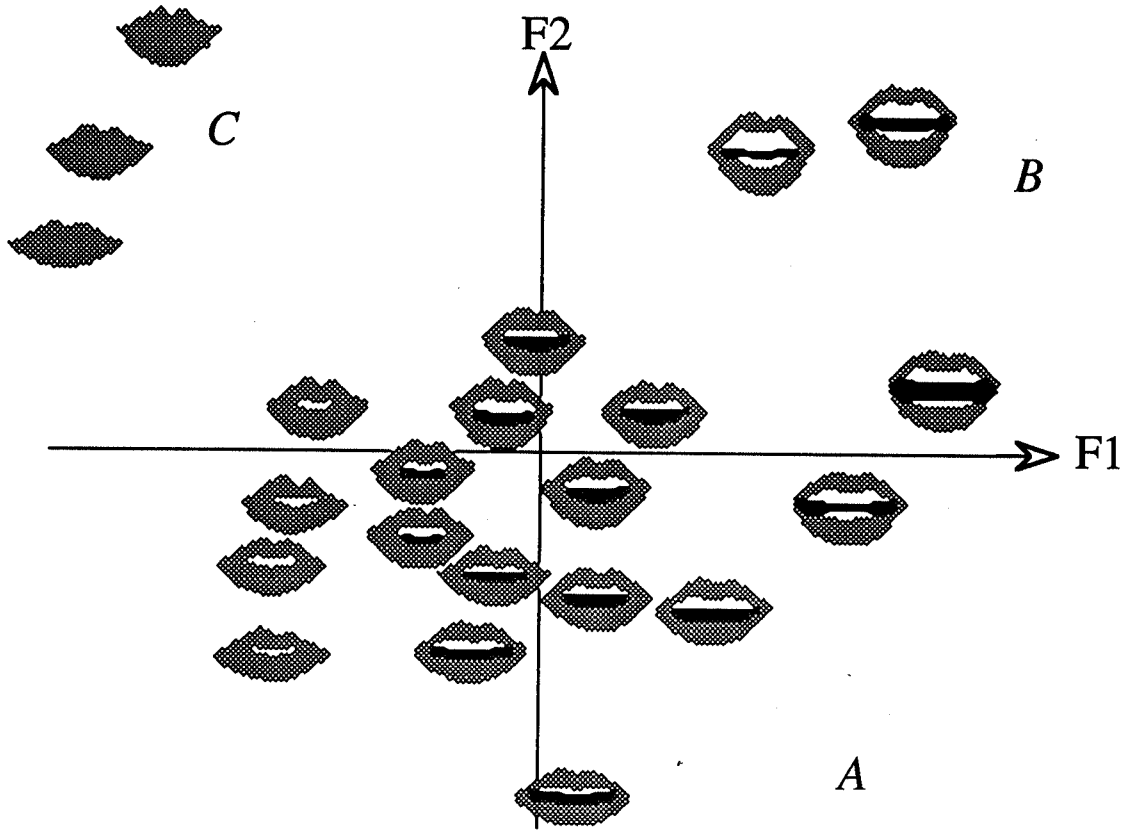


Figure 1.a. *Projection factorielle des 22 formes (schématisées) de lèvres et de dents utilisées pour la synthèse audio-visuelle de la parole. Les seules vues de face sont présentées ici (sans le menton) dans un plan essentiellement caractérisé par les jeux de la protrusion du contact labial (C), de l'étirement (A) et de la séparation (B) intérolabiales.*

2 . LE DICTIONNAIRE DE VISÈMES

Dans un étude récente (Benoît et al., 1992), il a été établi qu'une vingtaine de formes maxillo-labiales prototypiques d'un locuteur permettait de rendre compte de la géométrie des lèvres et du menton, au centre de la réalisation acoustique des 14 voyelles du français en isocontexte, des trois voyelles extrêmes [i, a, y] en différents contextes consonantiques, et des six consonnes [b, v, z, ʒ, ʀ, l] en différents contextes vocaliques. Les 23 classes – ou visèmes – révélées par cette analyse ne prétendent ni à une description fine de l'ensemble des formes maxillo-labiales significativement différentes, ni à une exhaustivité complète de tous les effets contextuels possibles en français. Pour des raisons évidentes, le corpus étudié était limité aux réalisations et aux contextes jouant un rôle majeur sur la dynamique des articulateurs visibles de la parole. Des règles fondamentales sur l'effet général des différents contextes, vocaliques et consonantiques, ont pu être observées à partir des données obtenues (Benoît et al., 1991), ouvrant la voie à leur *extrapolation* à l'ensemble des réalisations possibles. Enfin, les visèmes recensés correspondant aux formes caractéristiques au centre des réalisations acoustiques, une *interpolation* des formes est rendue nécessaire dans des conditions de débit lent où l'affichage d'une image intermédiaire est alors indispensable entre deux visèmes consécutifs : une grammaire des règles d'organisation temporelle de ces formes est également à l'étude en vue d'identifier, parmi les visèmes existants, ceux susceptibles d'être intercalés entre deux visèmes successifs, non contigüs dans l'espace multidimensionnel où ils ont été définis.

La figure 1 présente la projection factorielle (obtenue par analyse des correspondances sur l'ensemble de toutes les réalisations et des 14 paramètres géométriques mesurés sur celles-ci) des 22 visèmes caractéristiques de la parole (la position *repos – lèvres fermées* n'est pas représentée) et des trois paramètres les plus discriminants (A : écartement intérolabial ; B : séparation intérolabiale ; C : protrusion du contact labial). La figure 1.a présente la vue de face des lèvres et des dents correspondant aux réalisations des visèmes représentés par leur symbole à la figure 1.b. La liste par extension de ces mêmes visèmes est donnée dans le tableau 1. On notera, sur ces figures, l'important effet du contexte [ʒ] sur les voyelles [i] et [a], le long du facteur F2 (combinaision de protrusion et de séparation des lèvres), ou celui du contexte [z] sur [a], par exemple, qui *trans-forme* la forme du [a] "tenu" en celle d'un [i] "tenu" (par rapprochement des lèvres).

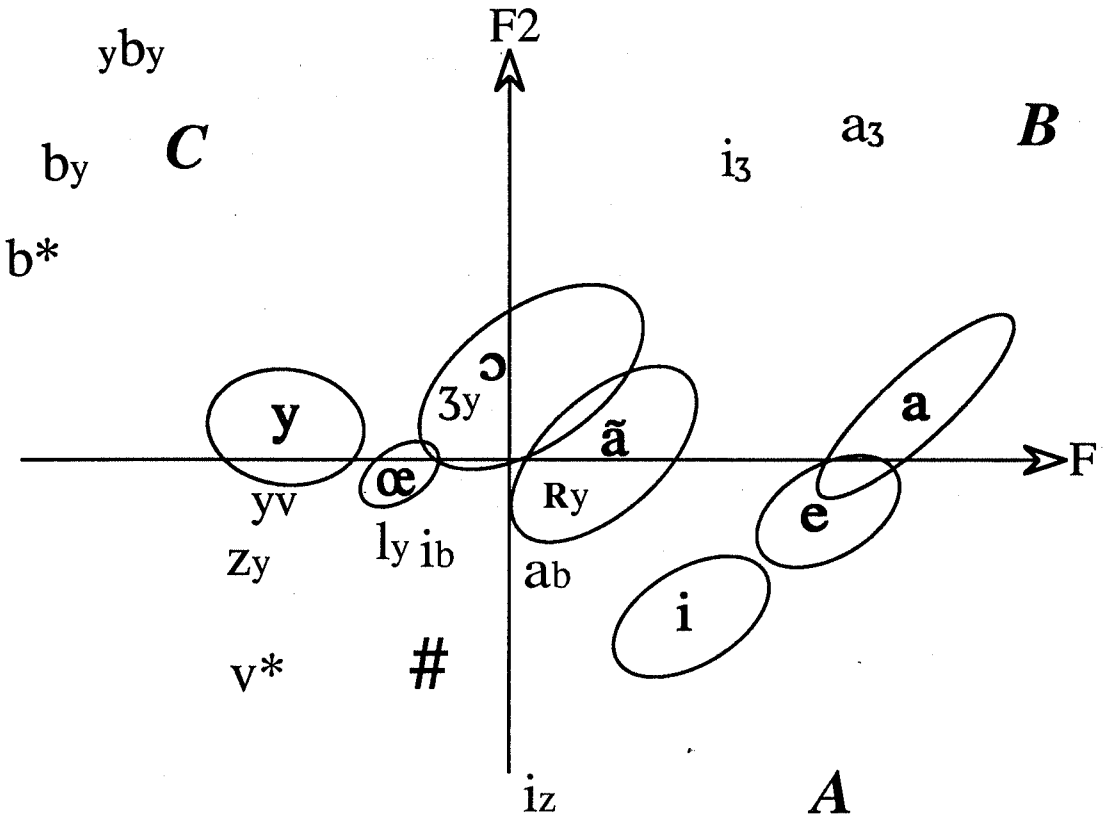


Figure 1.b. *Projection factorielle des 22 visèmes définis par leurs symboles (cf. leur définition dans le tableau 1. ci-dessous). Les ellipses de dispersion figurent les nuages de réalisations des visèmes vocaliques à 90 % de confiance.*

Tableau 1. *Liste des 23 visèmes avec leur symbole et leur contenu (visuo)phonétique majeur.*

Notation: * = [i] or [a]; V (c) = {[c$\underline{V}$c] / V ∈ [i, a, y]}; C (v) = {[v$\underline{C}$v] / v ∈ [i, a, y]};
C (v*) = {[v$\underline{C}$i], [v$\underline{C}$a] / v ∈ [i, a, y]}; C (*v) = {[i$\underline{C}$v], [a$\underline{C}$v] / v ∈ [i, a, y]}.

N°	**Symbole**	[$v_i\underline{V}_iv_iz$]	[$v_jc_k\underline{V}_ic_kv_jz$]	[$v_ic_kv_j\underline{C}_kv_iz$]
1	**{a}**	a	a (ʀ, l)	l (a), ʀ (a*, *a)
2	**{i}**	i	a (z)	
3	**{y}**	y, u, ø, o, õ	y (b, z, ʒ, ʀ, l)	[z, ʒ, ʀ, l] (y)
4	**{e}**	e, ɛ, ɛ̃	i (ʀ, l)	ʀ (i), l (i*, *i)
5	**{ɔ}**	ɔ		
6	**{œ}**	œ, œ̃		
7	**{ã}**	ã		
8	**{ab}**		a (b, v)	
9	**{a3}**		a (ʒ)	
10	**{ib}**		i (b, v)	
11	**{iz}**		i (z)	z (**)
12	**{i3}**		i (ʒ)	ʒ (**)
13	**{yv}**		y (v)	v (y)
14	**{b*}**			b (**)
15	**{by}**			b (y*, *y)
16	**{v*}**			v (**)
17	**{zy}**			[v, z] (y*, *y)
18	**{ʀy}**			ʀ (y*, *y)
19	**{ly}**			l (y*, *y)
20	**{ʒy}**			ʒ (y*, *y)
21	**{yby}**			b (y)
22	{#}	forme préphonatoire (lèvres entrouvertes)		
23	{=}	forme de repos (lèvres fermées)		

3. ARCHITECTURE DU SYSTÈME

3.a Structure générale du système

Le système que nous avons mis au point réalise une synthèse audio-visuelle de la parole à partir du texte par concaténation d'unités (acoustiques et optiques) préstockées. Comme le montre la figure 2., il utilise un dictionnaire d'environ 2000 polysons (avec sa table d'accès) sur lesquels sont réalisées des modifications prosodiques (par traitement PSOLA) d'une part, et une table d'accès à un dictionnaire de 23 visèmes nécessaire pour l'affichage final, d'autre part. Plusieurs dictionnaires peuvent être utilisés indifféremment (voix ou visage d'homme, de femme, etc.). L'entrée du système est un texte orthographique français quelconque ; les sorties sont un fichier d'échantillons audio sur 16 bits à 16 kHz et un fichier ASCII descripteur de la séquence d'images correspondante à 25 Hz.

L'ensemble des règles transformationelles nécessaires à la synthèse acoustique et visuelle à partir du texte (transcriptions orthographique-phonétique, phonétique-visémique, etc.), et des modélisations prosodiques ont été écrites sous le formalisme COMPOST (Bailly & Tran, 1989). Ce COMpilateur Phonétique sur Ordinateur pour la Synthèse de Texte permet la manipulation d'objets par des règles contextuelles. Ces objets sont organisés dans une structure arborescente dont ils sont les feuilles. Ils sont des instances d'objets génériques qui sont regroupés dans leur classe naturelle (mot, syllabe, phonème, archi-visème, visème, image, etc.) afin de les doter d'attributs numériques spécifiques.

Trois grammaires de règles ont été élaborées pour la conversion de l'arbre des phonèmes munis de leur durée (modélisée par le module prosodique de la synthèse acoustique) en une séquence d'images alignées sur le signal acoustique. Elles sont présentées ci-dessous.

3.b Correspondance phonème / archi-visème

Il s'agit d'une réécriture élémentaire. Par *archi-visème*, nous entendons ici un ensemble de configurations anatomiques identiques dans la réalisation d'un ou plusieurs phonèmes ne différant que par des traits articulatoires non visibles (voisement, nasalité, lieu de constriction dans le conduit vocal), et ce indépendamment du contexte. Ainsi, les occlusives bilabiales [p], [b], [m] se regroupent naturellement dans un seul archi-visème (noté V_P dans le tableau 2). La transcription se fait donc terme-à-terme, sans prise en compte du contexte, au seul niveau symbolique. L'archi-visème hérite de la durée du phonème correspondant. Les 14 archi-visèmes que nous utilisons actuellement pour le français sont présentés dans le tableau 2.

Tableau 2. Tableau de correspondance symbolique archi-visème – phonème.

archi-visème	phonèmes correspondants
V_A	a
V_E	e, ɛ, ɛ̃
V_I	i, j
V_U	y, u, ø, o, õ, w, ɥ
V_Œ	œ, œ̃
V_Ã	ã
V_Ω	ɔ
V_B	p, b, m
V_V	f, v
V_Z	s, z
V_J	ʃ, ʒ
V_R	ʀ
V_L	l
V_?	t, d, n, k, g, ɲ, ŋ

3.c Transcription archi-visème / visème

Pour nous, un visème est le code symbolique d'une classe de formes maxillo-labiales similaires regroupant divers allophones. Dans leur étude de nombreuses réalisations contextuelles, Benoît et al. (1992) ont identifié l'existence de 21 visèmes répondant à cette définition dans le jeu articulatoire d'un locuteur français représentatif du "français standard" prononcé sans expression faciale particulière, auxquels ont été ajoutés deux visèmes extra-linguistiques : la forme *préphonatoire*, aux lèvres entrouvertes ; et la forme *de repos*, aux lèvres jointes. La transcription archi-visème / visème utilise donc une grammaire de réécriture prenant en compte les contextes gauche et droit de l'archi-visème. Le visème transcrit hérite des propriétés de durée de l'archi-visème correspondant. Une séquence de visèmes est toujours initialisée par le visème *préphonatoire*, et conclue par le visème *de repos*. Un certain nombre de ces règles sont directement lisibles dans le tableau 1. Ainsi, l'archi-visème V_A entouré de deux archi-visèmes V_Z se transcrit sous la forme du visème noté {i} (cf. ligne 2, col. 4 du tableau 1), ou encore, l'archi-visème V_I dans un contexte V_R ou V_L se transcrit sous la forme du visème noté {e} (ligne 4, col. 4 ibidem). Nous soulignons par ces deux exemples la forte influence du contexte sur la forme maxillo-labiale d'une réalisation allophonique...

Toute la combinatoire contextuelle possible sur l'ensemble des archi-visèmes n'est pas implémentée dans la version actuelle du système présenté : elle est encore à l'étude.

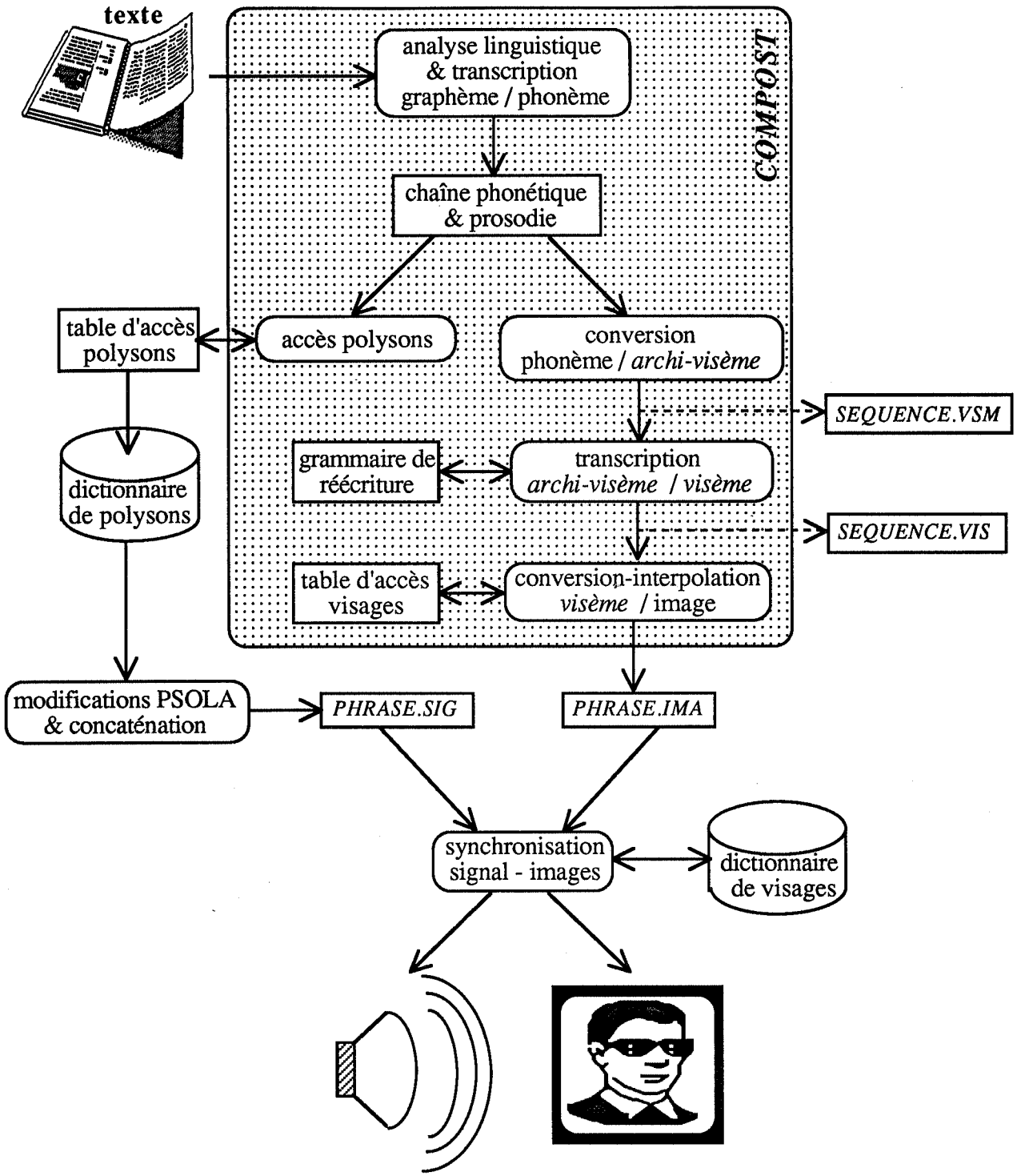


Figure 2. *Architecture du synthétiseur audio-visuel à partir du texte de l'ICP. L'environnement COMPOST est figuré dans le cadre pointillé. Le système fonctionne sur un PC doté de cartes OROS AU21 et VGA.*

3.d Transcription visème / image

Le dictionnaire d'images faciales utilisé par la synthèse doit naturellement être étiqueté de façon à ce que chaque image corresponde à (au moins) l'un des visèmes obtenus précédemment. Différents dictionnaires peuvent évidemment être utilisés : soit par simple modification de caractéristiques indépendantes des mouvements de la parole (nez, cheveux, etc.), soit par différents angles de vue (face, profil, etc.), soit enfin par la constitution d'un nouveau dictionnaire de formes avec un autre locuteur, à l'instar des dictionnaires d'unités acoustiques.

Cette dernière grammaire ne se limite pas au niveau symbolique. La transcription d'une chaîne de visèmes avec leur durée en séquence d'images étiquetées dans le dictionnaire utilisé doit tenir compte non seulement de la fréquence d'affichage des images sur l'écran (25 Hz dans notre cas pour des images 172 x 103 pixels), mais aussi de la possibilité d'interpolation spatiale d'une image entre deux images-cibles successives si la durée qui les sépare le permet. En effet, Le fichier de sortie contient une séquence d'images numérotées devant être visualisées à périodicité fixe par le synchronisateur son / image. Cette étape consiste donc à échantillonner les images nécessaires, en fonction de la durée prévue du visème courant et du décalage incrémental introduit par l'affichage des visèmes précédents (de durée constante). En outre, une matrice des interpolations possibles entre visèmes permet à ce stade d'intercaler, si cela est possible, l'image correspondant à un troisième visème entre deux (ou plus) images différentes. En plus clair, et sur l'exemple d'une transition entre les visèmes {i} et {a}, il est anatomiquement légitime d'intercaler l'image correspondant au visème {e} (voir figure 1.b) si la durée entre les images de {i} et de {a} est supérieure à 40 ms.

3.e Synchronisation image / son

Les deux étages de sortie du système de synthèse audio-visuelle à partir du texte fournissent un fichier acoustique binaire échantillonné à 16 kHz (PHRASE.SIG) d'une part, et un fichier ASCII descripteur de la séquence d'images à afficher à 25 Hz (PHRASE.IMA) d'autre part. Un programme *ad hoc* lit ces deux fichiers et les convertit de façon synchrone en signal audio via la carte OROS AU21 et en images animées sur la carte VGA respectivement. La commande d'affichage d'une nouvelle image est effectuée entre deux transferts du disque vers la mémoire du CNA, pendant le bref instant où l'unité centrale est disponible.

Un délai d'offset entre les deux canaux d'information image et son a été ajusté expérimentalement (à la trame près d'un enregistrement vidéo 50 Hz) de façon à ce que la première image "parole" soit synchronisée avec le début du signal audio. Ce retard est mis à profit pour afficher l'image correspondant à la forme préphonatoire {#}. En fin de séquence, la position de repos *lèvres fermées* reste affichée quelques instants.

6. CONCLUSION

Cet outil de synthèse de visage parlant à partir du texte avait été élaboré par Woodward en 1991 dans une première version. Il en est aujourd'hui à sa deuxième version et reste donc dans un état encore primitif.! Cependant, il a été conçu de façon à être évolutif. Il sera prochainement amélioré par une adjonction et une optimisation de règles de transcription. En outre, il est actuellement implanté sur un PC 386 ; et sa migration sur une machine plus puissante permettrait une visualisation de visages, soit plus grands et plus détaillés, soit plus fréquents. Enfin, l'intervention d'un infographiste permettrait de remplacer les rares et pauvres contours de notre visage artificiel par un dessin plus naturel – ou caricatural – qui lui fait actuellement défaut d'esthétique...

Mais la toute prochaine étape de notre travail consistera à quantifier l'apport d'intelligibilité de cette animation faciale à la parole synthétique par un test perceptif analogue à celui qui vient d'être réalisé avec de la parole et un visage naturels, et dont les résultats sont présentés par Mohamadi et Benoît dans ces mêmes Journées d'Etude sur la Parole.

Remerciements

Ce travail a été partiellement financé par l'ACCT. Nous remercions J. Zeiliger, A. Arnal, T. Lallouache et C. Abry pour leur précieuse collaboration.

Références

Bailly, G., & **Tran**, A. (1989), "Compost: A rule-compiler for speech synthesis", *Proc. 1st Eurospeech Conf.*, Paris, France, 136-139.

Benoît, C., **Boë**, L.J. & **Abry**, C.(1991), "The effect of context on labiality in French", *Proc. 2nd Eurospeech Conf.*, Gênes, Italie, 153-156.

Benoît, C., **Lallouache**, T., **Mohamadi**, T. & **Abry**, C.(1992), "A set of French visemes for visual Speech synthesis", in *Talking Machines : Theories, Models and Applications*, Bailly & Benoit Eds., Elsevier, North Holland, 485-504.

Binnie, C. A., **Montgomery**, A. A., & **Jackson**, P. L. (1974), "Auditory and visual contribution to the perception of consonants", *J. Speech & Hear. Res.*, 17, 608-618.

Erber, N.P. (1969), "Interaction of audition and vision in the recognition of speech stimuli", *J. Speech & Hear. Res.*, 12, 423-425.

Erber, N. P. (1975), "Auditory-visual perception of speech", *J. Speech & Hear. Dis.*, 40, 481-492.

Mohamadi, T., & **Benoît**, C. (1992), "Le gain des lèvres : intelligibilité auditive et visuelle de la parole bruitée en français", ce volume.

Neely, K. K. (1956), "Effect of visual factors on the intelligibility of speech", *J. Acoust. Soc. Am.*, 28, 1275-1277.

Sumby, W.H. & **Pollack**, I. (1954), "Visual contribution to speech intelligibility in noise", *J. Acoust. Soc. Am.*, 26, 212-215.

Summerfield, Q. (1979), "Use of visual information for phonetic perception". *Phonetica*, 36, 314-331.

Woodward, P. (1991), *Synthèse de visage parlant*, Mémoire de DEA Signal-Image-Parole, ENSERG, Grenoble, 43 p.

19es J.E.P. - Bruxelles, 19 au 22 mai 1992.

TREILLIS ACOUSTICO - PHONETIQUES : UNE METHODE D'EVALUATION

C. BOURJOT A. BOYER D. FOHR

CRIN, CNRS & INRIA LORRAINE
BP 239 54506 VANDOEUVRE CEDEX FRANCE

Résumé

Ce papier propose une méthodologie pour comparer le comportement de décodeurs acoustico-phonétiques. Traditionnellement, l'évaluation de ces systèmes consiste à calculer des matrices (de confusion, ...) à partir de chaînes linéaires phonétiques. Nous généralisons ces méthodes en travaillant sur les treillis phonétiques qui peuvent être pondérés ou non par des coefficients. Les matrices ne rendant plus compte de manière satisfaisante des résultats du décodeur, nous fournissons des courbes d'évolution des taux de reconnaissance en fonction du nombre de solutions retenues dans le treillis. La comparaison de décodeurs fournissant des treillis de taille très différente sera ainsi permise. Ce logiciel est entièrement paramétrable (définition des phonèmes, des macro-classes, de la hiérarchie des erreurs,...). Cela permet à l'utilisateur d'avoir ses propres critères d'évaluation.

INTRODUCTION

Un questionnaire distribué dans plusieurs laboratoires européens a montré la nécessité d'une évaluation quantitative et qualitative fiable des décodeurs acoustico-phonétiques (D.A.P.) [Bourjot 88].
Le but de cette évaluation est double :
- comparer deux systèmes différents ou deux versions d'un même système lors de la détermination du meilleur ensemble de paramètres de décodage.
- fournir des informations aux autres composants d'un système de reconnaissance de grand vocabulaire. En effet la connaissance des taux d'insertion de confusion et d'omission permet une utilisation plus efficace des sorties du D.A.P. par les autres modules du système de reconnaissance.

Dans nos précédents travaux [Bourjot 89a] nous avons étudié plusieurs algorithmes de programmation dynamique pour construire les matrices de confusion, d'insertion et d'omission nécessaires à l'évaluation. Ces algorithmes étaient adaptés à la fois aux matériaux de parole disponibles pour l'évaluation et au décodeur à évaluer, mais limités à l'évaluation de chaînes phonétiques linéaires. Rares sont les papiers [Thomson 89] qui traitent du délicat problème de l'évaluation de treillis.
Ce papier s'intéresse à ce problème et propose une méthodologie fondée sur la comparaison de courbes.

PROBLEME

Traditionnellement l'évaluation ne porte pas directement sur les performances du niveau phonétique mais traite de son influence sur les résultats globaux du système (au niveau du mot ou de la phrase). Ceci ne permet pas de connaître les performances intrinsèques de chacun des modules du système. Or, si on veut comparer deux systèmes qui utilisent des niveaux mots très différents (taille du lexique), la connaissance des taux globaux apporte peu d'informations. De plus, les treillis pouvant être de taille très différente, les taux ne sont pas comparables. Habituellement, la comparaison de deux décodeurs n'intégré pas la taille du treillis. Ainsi, cette comparaison se fera au détriment du décodeur qui fournit le treillis comportant le plus petit facteur de branchement. Nous proposons de tracer l'évolution des taux en fonction du nombre de chemins retenus dans le treillis.

SPECIFICATIONS

Nous considérons un treillis comme un graphe acyclique orienté éventuellement pondéré aux arcs et aux noeuds. Les systèmes fournissant simplement une suite de phonèmes constituent un cas particulier de treillis.

Notre système peut prendre en compte des coefficients de différente nature (plausibilité, probabilité, distance,...). Ces coefficients serviront à ordonner les chemins lors de la sélection d'un sous ensemble de solutions lors du tracé des courbes.

Les résultats fournis sont des courbes obtenues à partir de cinq taux : taux d'identité, de confusion légère, de confusion grave, d'insertion et d'omission. Une confusion légère est une confusion à l'intérieur d'une même macro-classe. Une confusion grave est une confusion entre deux phonèmes appartenant à des macro-classes différents. Un ensemble de macro-classes standard est fourni pour la langue française mais l'utilisateur a la possibilité de définir ses propres phonèmes et ses propres macro-classes. Le système est donc multi-lingue.

Les cinq taux ne suffisent pas à fournir une idée complète des performances du décodeur, car c'est leur prise en compte simultanée qui définit la qualité du treillis. C'est pourquoi il nous a paru intéressant de définir une combinaison linéaire de ces 5 taux. Une sixième courbe sera donc tracée, représentant l'évolution de cette mesure globale en fonction du nombre de chemins sélectionnés. Les coefficients standards de la combinaison linéaire sont fournis avec le logiciel, mais l'utilisateur peut définir ses propres coefficients.

METHODE

La méthode consiste à itérer sur n le traitement suivant :
- sélection de n chemins,
- calcul des 6 taux.

Ensuite, on trace les six courbes.

1- SELECTION DES CHEMINS

Un algorithme de type séparation et évaluation progressive (branch et bound) détermine les n meilleurs chemins dans le treillis. Si le décodeur ne fournit pas de pondérations sur les branches ou les noeuds du graphe, tous les chemins sont équiprobables et nous en sélectionnons aléatoirement n. Lorsque les coefficients sont des distances, les chemins sont ordonnés en fonction de la moyenne arithmétique des distances cumulées sur chaque chemin.
Le nombre de chemins retenus est un paramètre du logiciel. Si les deux décodeurs donnent des treillis de taille très différente, n est limité à la plus petite des deux valeurs.

2- CALCUL DES TAUX

La comparaison entre les chemins sélectionnés et une transcription à l'écoute de la phrase est réalisée par un algorithme de programmation dynamique. Les coefficients pour la combinaison linéaire des taux sont les mêmes que les pondérations utilisées par cet algorithme. Nous avons utilisé (1,1,1) pour la contrainte locale. Pour la distance locale, nous proposons:

- 100 pour une confusion grave
- 51 pour une insertion
- 51 pour une omission
- 10 pour une confusion légère
- 0 pour une identité.

Ces coefficients peuvent être modifiés par l'utilisateur qui a ainsi la possibilité de définir sa propre hiérarchie des erreurs.

Les taux sont calculés ainsi:
$[\Sigma_{i=1,ns} E(ARGMIN_{j=1,n} N(j))]/ ns$

n est le nombre de chemins retenus,
ns le nombre de phrases du corpus,
ARGMIN est la fonction qui fournit l'indice du minimum,
E(j) est le nombre d'erreurs du chemin j.
N(j) est le taux de dissemblance déterminé par programmation dynamique pour le chemin j.

Pour chaque phrase du corpus, on détermine dans un premier temps le chemin qui minimise le taux de dissemblance de programmation dynamique. Ensuite pour ce chemin, on comptabilise le nombre de phonèmes correctement étiquetés, le nombre de confusions, le nombre d'insertions, le nombre de confusions graves et le nombre de confusions légères. On moyenne ces taux pour toutes les phrases du corpus et on les affiche sous forme de courbes.

Comme il est impossible de distinguer une insertion suivie d'une omission d'une confusion grave, nous avons arbitrairement choisi de comptabiliser une confusion grave. C'est pourquoi les taux d'insertion et d'omission sont des estimations minorantes. De plus, comme nous ne disposons que de la transcription à l'écoute, il ne nous est pas possible de faire la distinction entre les séquences omission-confusion grave et confusion grave-omission. Cela devra être pris en compte lors de l'interprétation des courbes.

CORPUS

Pour évaluer un système de décodage acoustico-phonétique, on peut utiliser trois types d'étiquetage :
- un étiquetage aux frontières des phonèmes réalisés,
- une transcription à l'écoute,
- une transcription standard.

Lorsque l'on ne dispose que d'une transcription standard, on comptabilise de mauvaises prononciations comme des erreurs du système. Mais étiqueter aux frontières est une tâche fastidieuse et longue. Pour des corpus de taille

suffisante pour évaluer des systèmes, une transcription à l'écoute représente un bon compromis [BOURJOT 89b]. Le corpus est composé de phrases phonétiquement équilibrés (par exemple les phrases de Combescure pour le français). Le corpus doit avoir une taille suffisante pour fournir des résultats statistiquement valides. Trente phrases sont nécessaires puisque la différence d entre deux moyennes est significative si l'écart standard de la distribution des différences

$$Sd = RAC(s_1^2/n_1 + s_2^2/n_2)$$

est inférieur à 2 d pour un intervalle de confiance de 95%.
RAC est la racine carrée,
s_i l'écart type du système i,
n_i est le nombre de phrases dans le corpus.

INTERPRETATION DES COURBES

Le figure 1 donne l'exemple de courbes obtenues lors de la comparaison de deux versions d'un même décodeur. La figure 2 montre la comparaison de deux différents décodeurs. Ces courbes sont tracées tous phonèmes confondus.
Le nombre de chemins est en abscisse, le nombre moyen d'erreurs pour tout le corpus en ordonnée. De plus, en chaque point est représenté l'écart standard.

La figure 1 représente les courbes pour l'insertion, l'omission et la mesure globale pour le décodeur acoustico-phonétique APHODEX conçu au CRIN. On peut constater que les différents taux sont relativement constants parce que le nombre de chemins retenus (20) est relativement faible comparé au nombre de chemins possibles (environ 2000). Il est intéressant de constater que les deux versions sont équivalentes pour la mesure globale bien que les taux d'omission et d'insertion soient différents. Le résultat le plus important pour le concepteur du décodeur est de remarquer que lorsqu'il essaie de diminuer le taux d'omission, le taux d'insertion augmente.

La figure 2 représente la mesure globale pour deux décodeurs différents. L'abscisse est limitée à 180 car le décodeur fournit environ 180 chemins pour chaque phrase du corpus. Les performances des deux décodeurs ne s'améliorent plus au delà de 90 chemins. Le décodeur 1 est meilleur que le décodeur 2 à partir de 80 chemins.

CONCLUSION

Notre système étant entièrement paramétrable, un concepteur de systèmes de décodage peut utiliser ce logiciel pour déterminer les améliorations apportées par la nouvelle version de son système. Il peut définir les macro-classes qui correspondent le mieux à son problème et ainsi étudier l'évolution du décodeur pour une macro-classe donnée. En effet toutes les courbes peuvent être obtenues pour chaque macro-classe si la taille du corpus est suffisante.

Des résultats complémentaires peuvent être fournis. Par exemple, il est tout à fait possible de connaître le rang dans le treillis phonétique du chemin le plus proche de la phrase prononcée. La notion de proximité est bien sûr définie par l'évaluateur grâce aux paramètres de la programmation dynamique.

REFERENCES

[BOURJOT 88] : C. BOURJOT, A. BOYER, G. PERENNOU, N. VIGOUROUX, J.P. TUBACH, "Les systèmes à grand vocabulaire et leur évaluation, les systèmes de décodage phonétique et leur évaluation, état de l'art en Europe", 17èmes JEP, SFA-GCP, Nancy 1988.
[BOURJOT 89a] : C. BOURJOT, A. BOYER, D. FOHR, "Tools for phonetic labelling and phonetic assessment", Workshop on speech input/output assessment and speech databases, ESCA, Netherlands 1989..
[BOURJOT 89b] : C. BOURJOT, A. BOYER, D. FOHR, "Phonetic decoder assessment", EUROSPEECH, Paris 1989.
[THOMPSON 89] : THOMPSON H., "Evaluation of phoneme lattices : four methods compared", Workshop on speech input/output assessment and speech databases, ESCA, Netherlands 1989.

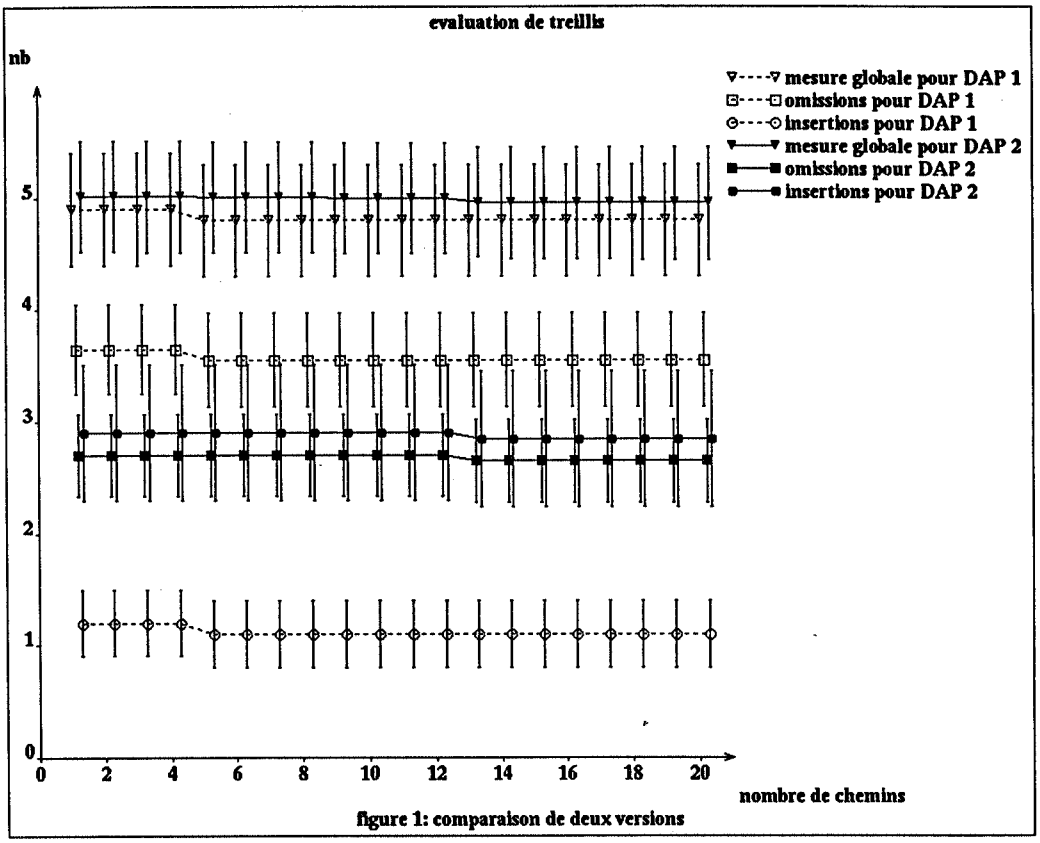


figure 1: comparaison de deux versions

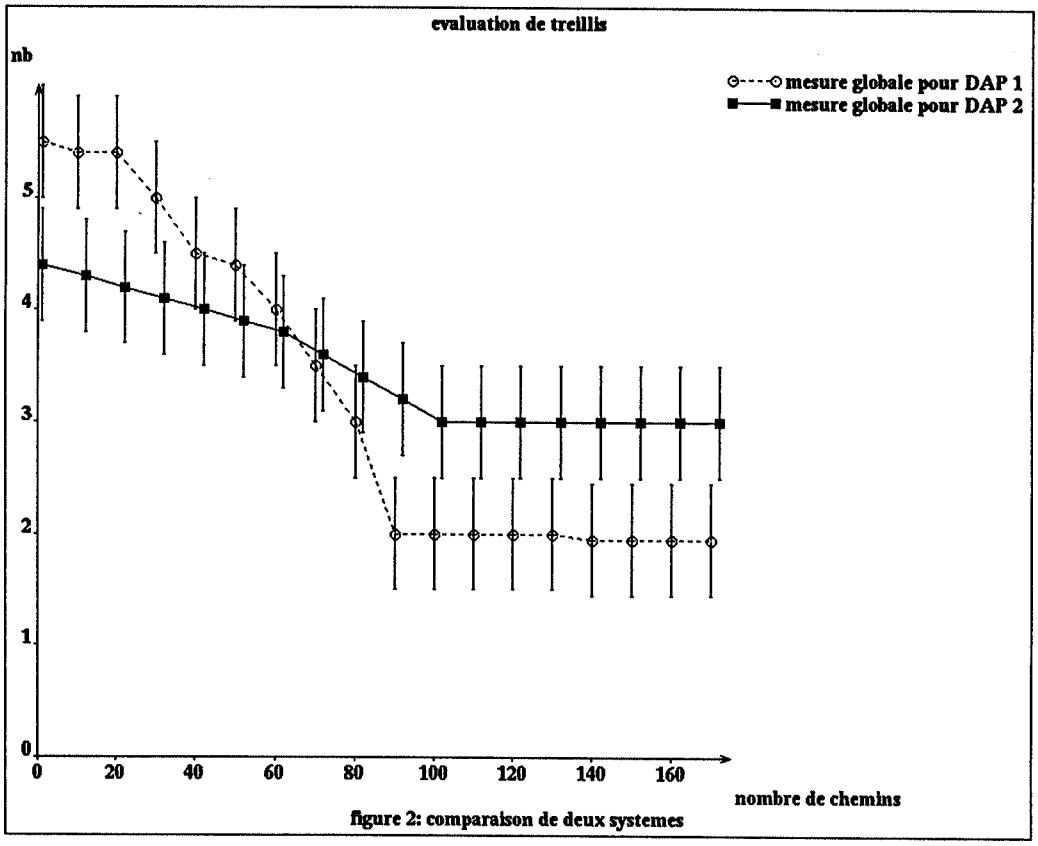


figure 2: comparaison de deux systemes

19es J.E.P. - Bruxelles, 19 au 22 mai 1992.

SYNTHESE A PARTIR DU TEXTE POUR LE CATALAN

J. CAMPS [1,2], G. BAILLY [1], J. MARTÍ [2]

(1) Institut de la Communication Parlée, URA CNRS n°368,
ENSERG/INPG-Université Stendhal. 46, av. Félix Viallet, 38031 Grenoble CEDEX, FRANCE.
(2) Ecole Universitaire de Télécommunication "La Salle". Université Ramon Llull.
Passeig Bonanova, 8. 08022 Barcelone, Catalogne, ESPAGNE.

Résumé

Nous présentons ici un système complet de synthèse à partir du texte mis au point pour la langue catalane. Ce système a été développé grâce aux outils multilingues offerts par le compilateur de règles COMPOST [Alissali et al., 91], en s'inspirant d'une maquette de synthèse du français.

Ce travail a consisté essentiellement a) à décrire la morpho-syntaxe du catalan b) à mettre au point un modèle de tri-grams, c) à réaliser l'accentuation, transcription et la syllabisation d'un texte quelconque, d) à définir et constituer un dictionnaire de polysons et enfin e) à écrire un jeu minimal de règles pour la génération prosodique.

Cet article présente la description et l'évaluation de l'état actuel de notre système.

1. INTRODUCTION

Ce travail se situe dans la perspective de conception d'une méthodologie de réalisation de systèmes de synthèse à partir du texte grâce à des outils multilingues. Nous présentons ici un système de synthèse du catalan réalisé suivant cette méthodologie. L'évaluation qui en est faite à la fin de cet article permet de conclure à la bonne performance de ces outils qui ont permis de créer de toutes pièces ce système temps-réel et évolutif en 12 mois de travail d'une seule personne.

2. DESCRIPTION MORPHO-SYNTAXIQUE

Nous avons réalisé la description morpho-syntaxique du catalan à l'aide d'un modèle de dérivation flexionnelle décrit par une grammaire à validation-saturation [Courtin, 77] : chaque morphème du lexique pointe sur un modèle. Le comportement morphologique de ce modèle est alors décrit par la grammaire par l'intermédiaire de règles qui valident ou saturent d'autres règles. Deux règles particulières déterminent les points d'entrée et de sortie de l'automate. Un modèle est validé si l'intersection entre l'ensemble des règles validées à l'état courant de l'automate et sa partie conditionnelle est non vide.

Le lexique morphologique contient donc les racines, préfixes, suffixes et désinences d'un grand nombre de mots qui potentiellement peuvent paraître dans le texte à synthétiser. Notre stratégie a consisté à décrire toutes les classes fermées du catalan (déterminants, conjonctions, prépositions...), les verbes usuels (à peu près 8000 racines verbales [Xuriguera, 85]), les homographes résultant de la description précédente ainsi que quelques adjectifs, adverbes et noms usuels. Le lexique dans son état actuel contient environ 17000 entrées.

La grammaire morphologique permet une description complète du catalan grâce à environ 280 modèles et 540 règles. Cette grammaire a été augmentée par un calcul de variables grammaticales (genre, nombre,...) qui ne sont pas pour l'instant exploitées.

Cette analyse lexicale nous permet d'émettre un certain nombre d'hypothèses en vue de leur filtrage syntaxique :lorsque l'automate peut atteindre son état de sortie à partir du caractère courant de l'entrée, il étiquette le mot avec toutes les classes lexicales calculées ; sinon, émet l'étiquette "classe inconnue" (réservées aux noms, adjectifs et adverbes qui constituent les classes ouvertes de la langue). Si un verbe non-usuel est employé, son étiquetage "classe inconnue" peut être remis en question par le filtrage syntaxique.

3. LE FILTRAGE : MODELE DE TRI-GRAMS

Préalablement à la transcription orthographique-phonétique, le texte passe par une étape de filtrage syntaxique qui est constitué par un modèle markovien. Ce filtre est un modèle de tri-grams [Derouault et al., 86] : les probabilités de transition sont données par des

histogrammes de comptage appris sur un corpus étiqueté à la main alors que les probabilités d'émission sont données par l'analyse lexicale. Afin de ménager l'hypothèse d'étiquetage lexical erroné ou incomplet, les classes lexicales non hypothétisées sont émises avec une probabilité faible mais non nulle. L'algorithme de Viterbi réalise donc ainsi un compromis entre étiquetage lexical et contraintes syntaxiques données par le modèle de tri-grams. Ceci permet de rendre compte des emplois non triviaux d'entrées lexicales (substantivations, emplois de verbes non usuels...).

4. LA TRANSCRIPTION ET LA SYLLABISATION

A niveau de la prononciation des phonèmes vocaliques, le catalan se montre hautement sensible à la position de l'accent dans les mots. Par exemple un graphème "O" tonique, sera transcrit toujours [o] ou [ɔ]. Par contre, si il est atone, et sauf quelques exceptions, il sera toujours transcrit [u].

Etant donné que le catalan est une langue à accent variable, la solution que nous avons retenu consiste à faire précéder la transcription proprement dite d'un examen du texte afin de détecter quelle est la voyelle tonique de chaque mot. La position de l'accent tonique est parfois indiqué graphiquement, mais dans la plupart des cas cette position doit être déterminée par des règles. A cet effet, deux grammaires contenant chacune une vingtaine de règles ont été écrites dans le formalisme COMPOST.

Avec les accents déjà déterminés, le texte accède a une nouvelle grammaire d'environ 450 règles de transcription orthographique-phonétique qui assure une correspondance entre les graphèmes et les allophones utilisés. A peu près la moitié de cet ensemble est employé pour la transcription des graphèmes "E" et "O".

Etant donné que les variations des réalisations vocaliques se réalisent sous l'accent tonique et que seulement les trois dernières syllabes des mots peuvent porter ce dernier, nous avons utilisé un dictionnaire inverse afin d'obtenir les règles de transcription et leurs exceptions [Bernal et al., 88].

Un troisième ensemble d'une trentaine de règles effectue la syllabisation de la chaîne phonétique.

5. DEFINITION DU DICTIONNAIRE DE POLYSONS

Nous avons conçu et créé un dictionnaire de polysons qui comprend 831 diphones et à peu près 700 tri- et quadriphones, en considérant les phonèmes /l,r,ɾ,j,w/ comme non-segmentables. En effet, il est généralement impossible d'identifier une partie stable dans les réalisations acoustiques de ces derniers phonèmes, particulièrement lorsqu'ils ne sont pas en position inter-vocalique.

Un jeu de 34 allophones ont été utilisés :/a, ə, e, ɛ, i, o, ɔ, u, j, w, p, t, k, b, d, g, β, ð, ɾ, l, ʎ, s, ʃ, z, ʒ, m, ɱ, n, ŋ, ɲ, r, ɾ, f, v/ auxquels il faut ajouter le silence /-/. Nous n'avons pas fait la distinction entre semi-voyelle et semi-consonne. De même, les réalisations /ʤ, ʧ, ʦ, ʣ/ n'ont pas été considérées comme des unités phonétiques mais comme un groupe consonantique.

Une matrice initiale représentant toutes les combinaisons entre sons nous a permis d'écarter celles qui phonotactiquement n'étaient pas possibles et de créer une liste de mots porteurs qui contenaient les paires qui potentiellement pouvaient apparaître dans une chaîne transcrite à partir du texte : à quelques exceptions près, les polysons sont donc extraits de mots de la langue sans avoir fait la distinction préalable sur la position de l'accent tonique.

5.1. Méthodologie d'enregistrement et d'étiquetage

Un premier corpus de polysons a été obtenu après l'enregistrement (10 Khz., 16 bits) et la segmentation semi-automatique [Wang et al, 90] d'un ensemble de 501 mots usuels qui contenaient les unités.

Un deuxième enregistrement a été nécessaire pour pouvoir corriger quelques omissions et erreurs de prononciation produites dans le premier. Ce deuxième corpus comprenait 82 mots.

Le corpus de mots étiquetés résultant a été ensuite réétiqueté afin de rendre compte de plusieurs problèmes de concaténation (mauvaise assimilation du voisement, relaxation des articulateurs en finale...). Ce réétiquetage a été réalisé en utilisant des diacritiques utilisés par l'algorithme d'optimisation du choix des unités par la synthèse par unités stockées. Les diacritiques utilisés sont les suivants :

X : Dévoisement des consonnes voisées,
Y : Voisement des consonnes non voisées,
Z : Autres problèmes : erreur du locuteur, réalisations peu claires...

Ce réétiquetage a affecté 309 réalisations phonémiques (240 modifiées avec X, 9 avec Y et 60 avec Z) dans 276 polysons.

Le symbole /q/ a été utilisé pour indiquer indifféremment le schwa final voisé et /qX/ pour indiquer la zone du burst des occlusives finales. Cet étiquetage a pour objectif de faciliter la segmentation par décomposition temporelle employée décrite plus bas.

5.2. Segmentation du corpus

La segmentation du corpus de mots porteurs a été faite à l'aide d'une méthode de décomposition temporelle [Atal, 83]. Cette méthode considère les caractéristiques spectrales du signal comme une combinaison linéaire de spectres cibles encadrant le spectre courant (centres des réalisations phonémiques). Les coefficients résultants de cette combinaison donnent une représentation du signal en termes de fonctions d'émergence.

La suite des fonctions d'émergence des différents réalisations d'un mot permet d'associer à chaque son sa

durée (frontières prises aux instants d'émergences égales), son centre de réalisation (centre de gravité de sa fonction d'émergence) et son noyau (zone stable déterminée par seuillage de la fonction d'émergence). Afin de permettre le contrôle prosodique du message synthétique, le dictionnaire de polysons contient, pour chaque son, l'ensemble de ces caractéristiques ainsi que la valeur de la fonction d'émergence à chaque période (utilisée par l'algorithme d'ajustement de l'énergie et de la durée).

6. REGLES PROSODIQUES ELEMENTAIRES

6.1. Détermination des patrons prosodiques

Tenant compte des informations de charge rythmique du message (nombre de syllabes, de phonèmes,..) véhiculées par la structure arborescente de COMPOST, le générateur prosodique mis au point permet de générer la description prosodique multiparamétrique (durée, intensité et fréquence fondamentale) de la phrase.

Bien que l'objectif de ce premier travail n'était pas l'étude en soi de la prosodie du catalan, une génération prosodique élémentaire est nécessaire à la création du signal de synthèse. Pour limiter le champ d'action, nous n'avons considéré que la modalité déclarative.

6.2. Modèle de génération

Le modèle retenu pour la génération de la mélodie [Fujisaki et al., 88] considère que la courbe mélodique est le résultat de l'addition de trois courbes :

- La ligne de base.
- Les patrons accentuels.
- .La micromélodie.

La ligne de base sert à décrire les mouvements mélodiques liées à la modalité de la phrase et au découpage de celle-ci en groupes de souffle. Elle permet de générer par exemple une ligne de déclinaison et l'intonème final. Les patrons accentuels sont appliqués pour l'instant sur chaque syllabe tonique dont le noyau est /i/ ou /u/. Un creux micromélodique d'amplitude caractéristique de chaque consonne est appliquée automatiquement. L'évolution de ces trois paramètres est décrite par des points-clés connectés par des trajectoires préétablies (splines, droites et trajectoires paraboliques).

Pour la durée et l'énergie, les valeurs intrinsèques de chaque phonème sont modifiés classiquement (par des facteurs multiplicatifs et additifs respectivement) par application d'un ensemble de règles qui tiennent compte de la nature tonique de la syllabe et du contexte phonétique immédiat. Par exemple, pour la détermination des durées des réalisations des phonèmes /j/ et /w/, la distinction est faite entre semi-voyelle et semi-consonne selon leur position dans la syllabe.

7. RECONSTRUCTION DU SIGNAL ACOUSTIQUE

Le signal de base obtenu après la concaténation de la suite de polysons integrés dans le message à synthétiser, doit être déformé pour l'adapter aux patrons proposés par le générateur prosodique.

Le contrôle de la durée et de la fréquence fondamentale est fait simultanément avec une technique PSOLA. Pour le contrôle de l'énergie, les valeurs contenues dans les consignes prosodiques sont modifiées en tenant compte de la valeur des fonctions d'émergence pour chaque instant. Notre dictionnaire de polysons contient donc le signal acoustique correspondant aux polysons sélectionnés et l'ensemble des marqueurs de périodes auxquelles sont attachés la longueur de la période, son indice de voisement ainsi que la valeur en début de période de la fonction d'émergence associée à la cible immédiatement antérieure.

7.1. Contrôle de la durée et de la mélodie

L'utilisation préalable d'un algorithme de marquage des périodes [Barbe et al., 90] nous permet de réaliser la déformation prosodique des signaux originaux période par période grâce à une technique TD-PSOLA [Moulines et al., 89]. Notre algorithme utilise une fenêtre de Hanning asymétrique (figure .1).

A l'aide d'une courbe de remplissage qui assure que les variations les plus importantes du signal vont se produire dans le noyau de chaque réalisation phonémique [Bailly et al., à paraître], on fait correspondre à chaque période de synthèse une période du signal d'analyse. Cette technique est donc pilotée par la synthèse. La demi-longueur de la fenêtre de Hanning Th utilisée pour déformer une période de longueur Ta en une période de longueur Ts est égale au minimum des deux précédentes : Th = min(Ts, Ta).

Avec cette méthode, le signal au voisinage de l'instant de fermeture glottique (pendant laquelle il y a une excitation maximale du conduit vocale) est recopiée intacte dans les périodes de synthèse. Il est d'autre part à souligner qu'en cas d'absence de modification prosodique, le signal original est recopié sans distorsions.

7.2. Contrôle de l'énergie

A un lissage des contours d'énergie aux jonctures inter-polysons [Rodet et al., 90], nous avons préféré un contrôle explicite de l'energie par la donnée de sa valeur au noyau de chaque réalisation phonémique. Ainsi, l'écart entre la valeur de la période centrale du noyau du signal original et la valeur exigée par le générateur prosodique est calculé.

C'est la fonction d'émergence qui réalise l'interpolation entre les écarts constatés à des noyaux successifs. Cette stratégie permet donc un lissage de la courbe d'energie car elle est appliquée à gauche et à droite de la période centrale du noyau qui est précisément le lieu de découpage des polysons.

8. EVALUATION

L'intelligibilité au niveau mot du système a été évaluée. Cette expérience a consisté à faire écouter dans une salle de classe de l'Université Ramon Llull une suite phonétiquement équilibrée de 50 mots [Pialloux et al., 78] à 30 étudiants naïfs dans un choix ouvert. Pour se familiariser avec la synthèse, nous avons fait écouter trois phrases déclaratives. Les stimuli ont été entendus par les auditeurs qu'une fois. Des 1500 réponses, 1291 (86,1 %) sont correctes, 168 (11,2 %) ont été percues incorrectement et les 41 (2,7 %) restantes n'ont pas été identifiées comme des mots de la langue (cf. Table.1). Il est à noter le fait que 66 des réponses non correctes se concentrent sur trois mots seulement ("HOSTE", "ONCLE", "BROCA"). Ces erreurs ,dues principalement à des erreurs d'identification d'occlusives, peuvent s'expliquer par un modèle de génération de durée d'occlusion inapproprié. La majorité des autres confusions ont été en faveur de mots plus communs.

Cette première évaluation démontre l'apport indiscutable de la synthèse temporelle vis à vis de la synthèse par formants, puisqu'en utilisant le même corpus de mots et la même méthodologie, le système de référence [Martí, 86] atteint seulement 81,5 % d'intelligibilité.

9. CONCLUSIONS ET PERSPECTIVES

Le système décrit plus haut permet d'effectuer la synthèse en temps-réel d'un texte quelconque en catalan dans une bonne qualité. Les diverses étapes de traitement ont été entièrement décrites dans le formalisme de l'environnement de programmation COMPOST. Ceci permet d'envisager une évolution progressive du système vers une synthèse de plus haute qualité en améliorant les divers modules de traitement, notamment le module de génération prosodique.

Le dictionnaire de polysons aurait pu bénéficier d'un étiquetage plus soigné des réalisations, notamment par une distinction entre voyelles toniques et atones, et par l'étude plus approfondie des segments transitoires.

Nous envisageons de faire une étude assez complète des problèmes de concaténation afin d'en évaluer les conséquences perceptives et de proposer une représentation du signal et une méthode d'interpolation plus apte à pallier les problèmes d'étiquetage résiduels.

REMERCIEMENTS

Nous remercions M. Alissali pour sa précieuse collaboration. I. Roca et M. Moneo nous ont grandement aidés dans nos combats avec la phonétique catalane.

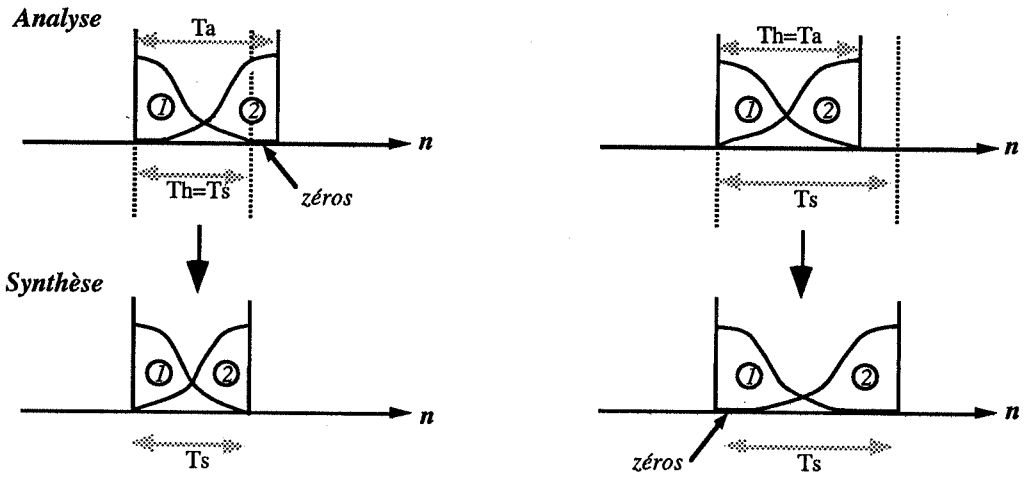


Figure .1 : Application de la technique TD-PSOLA utilisant une fenêtre asymétrique.A gauche est figuré le glissement de fenêtres permettant une diminution de la période du fondamental et à droite le glissement pour une augmentation.

REFERENCES

Atal. S.B. (1983) : "Efficient coding of LPC parameters by temporal decomposition". *IEEE Conference on Acoustics Speech and Signal Processing.* Vol.1. pp 81-84.

Alissali, M., Bailly, G. (1991) : "COMPOST : Un serveur de synthèse multilingue". *8ème. Congrès Reconnaissance des Formes et Intelligence Artificielle.* AFCET. Lyon. pp. 183-193.

Bailly, G., Barbe, T., Wang, H. (à paraître) : "Automatic labelling of large prosodic databases : Tools, methodology, and links with a text-to-speech system". *Talking Machines : Theories, Models and Designs.* Bailly et Benoît, Ed. North-Holland.

Barbe, T., Bailly, G.(1990) : " Evaluation d'un détecteur de fréquence fondamentale d'un signal microphonique par comparaison avec la mesure effectuée sur le signal laryngographique", *18èmes Journées d'Etudes sur la Parole,* Société Française d'Acoustique, pp. 165-169.

Bernal, M.C., Codina, F, Fargas, A, Tió, J. (1988) : "Diccionari invers : Llista inversa de les entrades del Primer Diccionari". *Quaderns de Didàctica de la Llengua.* E.U. Vic. Ed. Eumo. Barcelone.

Courtin, J. (1977) : "Algorithmes pour le traitement intéractif des langues naturelles". *Thèse de Doctorat. Université Scientifique et Médicale de Grenoble. INPG.* Grenoble.

Derouault, A.M, Mérialdo, B.(1986) : "Natural language modeling for phoneme to text transcription". *IEEE Transactions on Pattern Analysis and Machine Intelligence.* vol. PAMI-8. n.6. nov.1986.

Fujisaki, H., Kawai, H. (1988) : "Realization of linguistic information in the voice fundamental frequency contour of the spoken japanese". *Proc. IEEE Int. Conf. Acoust, Speech, Signal Processing.* pp. 663-666.

Martí, J. (1986) : "Un conversor text-veu en català : Sistema SINCAT (SINtetitzador en CATalà). *Actes du 3ème. Congrès de Langages Naturels et Langages Formels.* Université de Barcelone. C. Martín, Ed.

Moulines, E., Charpentier, F., Hamon, C. (1989) : "A diphone synthesis based on time-domain prosodic modifications of speech".. *Proc. IEEE Int. Conf. Acoust, Speech, Signal Processing.* pp. 238-241.

Pialloux, J., Valtat, M., Freyss, G., Legent, F.(1978): "Manual de Logopèdia". Traduction et adaptation de J. Perelló. Ed. Masson - Toray. Barcelone.

Rodet, X., Depalle, Ph., Poirot, G. (1990) : " Energy and articulation for improving diphone speech synthesis". *Proc. of the Workshop on Speech Synthesis.* Autrans (France), 25-9-90.

Wang, H.D., Bailly, G., Tufelli, D. (1990) : "Automatic segmentation and alignement of continuous speech based on the temporal decomposititon model". *International Conference on Spoken Language Processing,* Kobe, pp. 457-460.

Xuriguera, J.B (1985) : "Els verbs catalans conjugats". *Col. Pompeu Fabra, n.6.* Ed Claret. Barcelone.

Table.1. Résultats du test d'intelligibilité (PR signifie pas de réponse).

mots prononcés	confusions (nombres d'occurence)							nombre total d'erreurs ↘
suro	xurro (1)	xuru (1)						2
ànec								0
figa	fira (8)	fina (1)						9
luxe	fluxe (2)	boxa (1)	polze (1)	cursa (1)				5
bitlla	vella (1)							1
pruna	una (1)							1
vidre	vida (9)							9
cèrvol	tèrbola (1)	cèl-lula (1)						2
metge								0
truja	pluja (3)	PR (1)						4
ungla	poblat (8)	un guant(1)	poble (1)	umbral (1)	PR (6)			17
marbre	marulla (1)	PR (2)						3
fetge								0
canya	gaire (1)	calla (3)						4
ferro	zero (4)	gerro (1)	PR (1)					6
joia								0
tendre	tèbia (1)	membre (1)	PR (1)					3
xiscle	xispa (2)							2
polze								0
llàntia								0
gorja	borja (2)	boja (1)	glòria (1)	PR (4)				8
oncle	omple (12)	pompa (1)	ompat (1)	PR (7)				21
dutxa								0
xàfec								0
terme								0
illa	villa (1)	filla (3)	guilla (1)					5
hoste	pasta (18)	costa (1)	posta (4)	postres (1)				24
bruixa	maduixa (1)							1
murri	PR (2)							2
petja	metge (14)	fetge (3)	PR (2)					19
fàstic								0
rauxa	drauxa (3)	grauxa (1)						4
sutge	jutge (2)							2
borni	PR (2)							2
euga	heure (1)							1
gendre	llendra (1)	PR (3)						4
poma	polmer (1)							1
llavi								0
branca	llarga (1)	branga (2)	PR (3)					6
cel-la	PR (1)							1
fusta								0
enze	entra (3)	ensla (1)	quinze (3)	herba (1)	setze (1)	esa (1)	PR (6)	16
jeure	lleure (1)							1
avi								0
broca	droga (12)	groga (7)	broga (2)					21
bleda								0
índex								0
molla								0
piga	biga (2)							2
lletja								0

19es J.E.P. - Bruxelles, 19 au 22 mai 1992.

DECOMPOSITION TEMPORELLE ET RUPTURES DE MODELES POUR LE DECODAGE ACOUSTICO-PHONETIQUE

C. BARRAS[1], M.-J. CARATY[1], P. DELEGLISE[2], C. MONTACIE[1], R. ANDRE-OBRECHT[3] & X. RODET[1]

[1]LAFORIA - Université Paris 6, CNRS-URA 1095, 4, place Jussieu, 75252 Paris Cedex 5
[2]ENST, Dépt. SIGNAL, CNRS-URA 820, 46 rue Louis Barrault, 75634 Paris Cedex 13
[3]IRISA - Centre INRIA de Rennes, Campus de Beaulieu, 35042 Rennes Cedex

Résumé

Nous proposons dans cet article une amélioration du décodage acoustico-phonétique au moyen de la coopération de plusieurs méthodes d'analyse. Nous avons étudié deux méthodes a priori complémentaires : la Décomposition Temporelle et les Ruptures de Modèles. La Décomposition Temporelle modélise la trajectoire spectrale du signal de parole, tandis que les Ruptures de Modèles détectent les discontinuités du signal. Chaque méthode fournit sa propre segmentation du signal de parole. Ces segmentations ont été évaluées et comparées d'après des expériences de décodage acoustico-phonétique. En utilisant une paramétrisation MFCC du signal et un alignement dynamique entre les segments, chacune des méthodes donne un taux d'identification d'environ 80%, mais présente un taux élevé d'insertion (supérieur à 25%). La coopération des deux méthodes dans un système de décodage nous a permis d'obtenir un taux d'identification de 72% pour 7% d'insertion, et ainsi de corriger une grande partie des erreurs commises par chacune des méthodes.

1. INTRODUCTION

Bien que les performances des systèmes actuels de reconnaissance de la parole soient en progression, une avancée notoire n'est possible que par une amélioration importante du décodage acoustico-phonétique (DAP), car aucune des diverses méthodes proposées pour cette étape de décodage ne donne entière satisfaction. Cependant, certaines nous semblent complémentaires, et leur coopération devrait permettre de dépasser les limites actuelles du DAP.

C'est pourquoi nous proposons suivant cette approche une première étude conjointe de la Décomposition Temporelle (DT) [1] et des Ruptures de Modèles (RM) [2]. Ces deux méthodes segmentent le signal de parole selon deux principes différents : la première méthode (DT) modélise l'évolution spectrale et permet la localisation de cibles acoustiques ; la seconde (RM) détecte les discontinuités du signal de nature également spectrale, afin de faire apparaître des zones transitoires et "quasi-stationnaires". Cette double analyse est donc particulièrement bien adaptée au signal vocal, dont on connait toute la complexité.

Nous avons tout d'abord étudié les résultats respectifs des deux méthodes pour l'identification phonétique. Cette analyse a mis en évidence les taux d'insertion élevés de ces deux méthodes. Pour diminuer le taux d'insertion, nous avons intégré les deux méthodes en exploitant leur complémentarité. Cela a permis une augmentation du taux global de décodage et de sa fiabilité.

2. METHODES D'ANALYSE ET DE SEGMENTATION

Deux méthodes d'analyse du signal ont été étudiées. La Décomposition Temporelle étudie les déformations de la trajectoire spectrale pour localiser les principaux événements du signal de parole. Les Ruptures de Modèles sont fondées sur une étude statistique du signal, modélisant une suite de zones quasi-stationnaires par un modèle auto-régressif. Chacune de ces méthodes permet de définir une segmentation du signal de parole.

2.1. Décomposition Temporelle

La Décomposition Temporelle [1] décrit l'évolution spectrale du signal de parole par un modèle d'interpolation linéaire. Soit $\{y_n\}_{n=1..N}$ une suite de N vecteurs spectraux de dimension p, l'estimation y_n^* du vecteur spectral y_n est la combinaison linéaire de m spectres discrets, les cibles $\{g_k\}_{k=1..m}$, pondérés par des fonctions d'interpolation compactes $\{\phi_k(n)\}_{k=1..m, n=1..N}$. Ces fonctions d'interpolation représentent le support temporel de chaque événement acoustique et leur cible associée représente sa caractéristique spectrale.

Ainsi, l'estimation y_n^* du vecteur y_n est la suivante :

$$y_n^* = \sum_{k=1}^{m} g_k \phi_k(n)$$

L'algorithme DT optimise le choix du nombre d'événements n, des vecteurs spectraux $\{g_k\}$ et des fonctions d'interpolation $\{\phi_k(n)\}$. Le critère d'optimalité dépend de l'application choisie (e.g., codage[1], synthèse[3], reconnaissance[4]). Pour le DAP, il est souhaitable que $\phi_k(n)$ représente l'influence du $k^{ième}$ phonème dans la trajectoire spectrale, et que le vecteur g_k correspondant représente ce phonème dans l'espace des paramètres. En raison de la non-linéarité des transitions dans cet espace et des phénomènes asynchrones lors de la réalisation d'un phonème, des cibles appelées cibles de transition sont insérées. L'algorithme utilisé doit minimiser le nombre de cibles de transition tout en détectant le plus de phonèmes possible.

L'estimation des fonctions d'interpolation $\{\phi_k(n)\}$ se fait en détectant les déviations significatives de la trajectoire spectrale de $\{y_n\}$. Pour trouver ces déviations, on cherche les projections représentatives de la trajectoire spectrale qui sont les plus similaires à une famille de fonctions $\{\omega(n)\}$ (cf. Figures 1 & 2). La mesure de similarité M_S utilisée [5] entre les fonctions $\{\phi_k(n)\}$ et $\{\omega(n)\}$ est la suivante :

$$M_S(\phi,\omega) = \min_{c \in R} \left(\frac{\sum_{n=n_1}^{n_2} (\phi(n)-c)^2 . \omega(n)^2}{\sum_{n=n_1}^{n_2} (\phi(n)-c)^2 \cdot \sum_{n=n_1}^{n_2} \omega(n)^2} \right)$$

Le choix de la famille de fonctions et la sélection des fonctions d'interpolation sont fondés sur l'utilisation d'un algorithme de fenêtrage adaptatif [4].

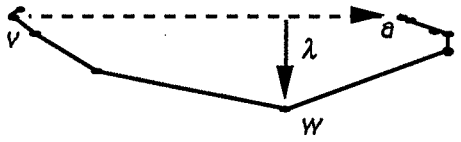


Figure 1. Trajectoire paramétrique dans le plan principal d'inertie du segment de parole [vwa]

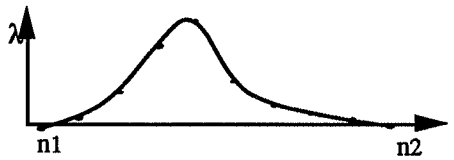


Figure 2. Projection de la trajectoire sur l'axe λ en fonction du temps

Les supports temporels des fonctions d'interpolation sur le continuum vocal, pouvant présenter des zones de recouvrement, constituent la segmentation issue de la technique de Décomposition Temporelle.

2.2. Ruptures de Modèles

La modélisation par Ruptures de Modèles, plus communément appelée méthode de Divergence Forward-Backward, est fondée sur une étude statistique du signal. Elle a pour caractéristiques :

- un calcul séquentiel des paramètres à chaque échantillon,
- l'application d'un test statistique (modèle-test),
- une détection séquentielle des ruptures.

Le signal $\{s_n\}$ est supposé être décrit par une suite de zones quasi-stationnaires caractérisées par un modèle auto-régressif gaussien d'ordre p :

$$s_n = \sum_{i=1}^{p} a_i s_{n-i} + e_n$$

où $\{e_n\}$ est un bruit blanc gaussien de variance σ_n^2.
En supposant que l'écart-type σ_n est constant le long d'une zone quasi-stationnaire, le modèle est défini par le vecteur :

$$(A^T, \sigma) = (a_1, \ldots, a_p, \sigma), \text{ noté } M(A, \sigma)$$

La détection des ruptures revient à détecter les changements du vecteur $M(A, \sigma)$.
Dans ce but, deux modèles $M_0(A, \sigma)$ et $M_1(A, \sigma)$ sont identifiés à chaque instant sur deux fenêtres d'analyse distinctes : une fenêtre croissante et une fenêtre glissante (cf. Figure 3).

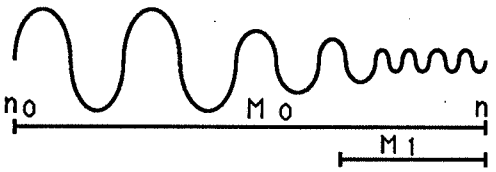


Figure 3. Fenêtres d'identification des modèles M_0 et M_1

Si aucun changement dans les paramètres n'intervient entre l'instant n_0 (i.e., instant correspondant à la dernière rupture détectée) et l'instant courant n, les modèles M_0 et M_1 restent identiques et toute distance définie entre eux doit être nulle. La distance statistique utilisée est obtenue à partir de l'entropie mutuelle $\{w_n\}$ des deux lois conditionnelles correspondantes calculée à chaque instant. Le test statistique est la somme de ces valeurs cumulées sur l'intervalle de temps. Une étude statistique [2] montre que la recherche des ruptures revient à détecter les changements de moyenne. On utilise, à cet effet, la règle de Page-Hinckley : à chaque valeur w_n est ajoutée un biais constant δ, et il suffit alors de détecter les maxima locaux supérieurs à un certain seuil λ.
Soit :

$$\hat{W}_n = \sum_{k=1}^{n} (w_k + \delta)$$

avec :

$$w_k = \frac{1}{2}\left[2\frac{e_k^0 e_k^1}{\sigma_1^2} - \left[1 + \frac{\sigma_0^2}{\sigma_1^2}\right]\frac{e_k^0}{\sigma_0^2} + \left[1 - \frac{\sigma_0^2}{\sigma_1^2}\right]\right]$$

et :

$$e_k^i = s_k - \sum_{j=1}^{p} a_j^i s_{k-j}, \quad i=0,1$$

où e_k^i est l'erreur de prédiction à l'instant k pour chacun des modèles M_i (A_i,σ_i), (i=0,1).
Une rupture est détectée à l'instant n_D défini par :

$$n_D = \underset{n \geq 0}{\text{Argmin}} \left\{ \max_{1 \leq k \leq n} (\hat{W}_k - \hat{W}_n) > \lambda \right\}$$

et la rupture a lieu à l'instant r :

$$r = \text{Argmax}_{1 \leq k \leq n_D} \{ \hat{W}_k \}$$

La mise en œuvre de cette méthode nécessite le choix des paramètres suivants :
- l'ordre p des modèles M_0 et M_1 ,
- la longueur de la phase d'initialisation (i.e., la période critique durant laquelle la recherche de ruptures est interrompue),
- le biais δ , et le seuil λ .

Ces paramètres ont été appris sur une phrase extraite du corpus de phrases phonétiquement équilibrées PEQ de BDSONS. L'ordre p des modèles est fixé à 2, la période d'initialisation à 20 ms. Il est apparu pertinent de distinguer deux zones pour fixer le biais δ et le seuil λ : les zones voisées et les zones non-voisées. Deux couples ont été introduits pour chacune des zones :
- (δ_v, λ_v) = (0.2 , 40) pour les zones voisées,
- $(\delta_{nv}, \lambda_{nv})$ = (0.8 , 80) pour les zones non-voisées.

Le test de voisement est effectué pendant la phase d'initialisation.
Les premières expériences ont montré que certaines frontières étaient systématiquement omises (e.g., la frontière [e m] est omise contrairement à celle [m e]). Ces résultats expérimentaux et une étude théorique [2] ont conduit à calculer la statistique dans le sens rétrograde du signal pour tout segment estimé trop long. Cette procédure complète, appelée méthode de Divergence Forward-Backward, est l'algorithme utilisé pour cette étude.
Les segments du continuum vocal bornés par deux ruptures consécutives constituent la segmentation issue de la technique des Ruptures de Modèles.

3. BASE DE DONNEES ET PARAMETRISATION

3.1. Base de données

Les expériences ont été effectuées sur les corpus SYLy-Acoustique de la base de données BDSONS enregistrés par deux locuteurs : un locuteur masculin (BP) et un locuteur féminin (MD). Ces corpus, numérotés (y) de 1 à 12, permettent l'étude des 192 diphones $[C_nV_y]$ dénombrés pour les voyelles orales ou nasales $[V_y]$ (y=1,..., 12) et les consonnes $[C_n]$ (n=1,..., 16) du français. Un corpus SYLy est constitué de 16 phrases $\{P_n\}$ (n=1,..., 16), une phrase P_n contenant plusieurs occurrences du diphone $[C_nV_y]$.
La base de données comprend plus de 4000 phonèmes répartis en 28 classes phonétiques, dont 12 voyelles (i.e., [a], [i], [e], [ɛ], [y], [ø], [u], [o], [ɔ], [ɑ̃], [ɛ̃], [ɔ̃]) et 16 consonnes (i.e., [p], [t], [k], [b], [d], [g], [m], [n], [f], [s], [ʃ], [v], [z], [ʒ], [l], [ʀ]).
Les signaux de parole ont été étiquetés manuellement sur le principe de l'étiquetage large [6].

3.2. Paramétrisation

Le signal est représenté par une suite de vecteurs de 10 coefficients MFCC (Mel Frequency Cepstral Coefficients). Ces coefficients sont calculés à partir des énergies extraites de 24 filtres triangulaires [7] répartis sur l'échelle de fréquence Mel [8]. Le spectre est calculé par une technique de lissage cepstral itératif [9] qui estime l'enveloppe à partir des pics stables. L'amélioration de cette technique par rapport aux méthodes classiques est surtout sensible dans le cas des voix féminines.

3.3. Segmentation

La segmentation issue de DT comporte plus de segments que de phonèmes réellement présents (12892 cibles pour 8220 phonèmes, soit 57% de segments supplémentaires). Les segments obtenus par RM sont beaucoup plus nombreux (17766) car les ruptures localisent les événements transitoires (e.g., les explosions, le début et la fin de voisement).
Avant même d'avoir effectué les expériences de DAP, nous savons qu'il faut un traitement spécifique des plosives, puisque la méthode RM détecte une rupture sur l'explosion des plosives. Il est donc nécessaire de caractériser de telles ruptures avant d'identifier la plosive sur le segment approprié. Ce traitement est actuellement en cours de réalisation.

4. EVALUATION DES SEGMENTATIONS

Deux segmentations distinctes des signaux sont obtenues à partir des méthodes DT et RM. Pour une segmentation donnée, nous définissons des segments-référence relativement à l'étiquetage manuel. Ces segments-référence sont les segments contenant une étiquette unique de l'étiquetage large. Ces segments a priori les plus porteurs de l'information phonétique serviront de référence pour la phase d'apprentissage.

Les segmentations sont évaluées par la cohérence de l'ensemble des segments-référence et par le DAP de l'ensemble des segments.
Deux distances inter-segments sont testées :
- la distance euclidienne $D_{Moy}(S_r,S_t)$ entre les vecteurs MFCC moyens des segments S_r et S_t.
- la comparaison dynamique $D_{Dyn}(S_r,S_t)$ entre les

segments, fondée sur la distance euclidienne des vecteurs MFCC.

4.1. Cohérence de l'ensemble des segments-référence

Chaque segment de l'ensemble des segments-référence est comparé aux autres. La liste des étiquettes des segments les plus proches au sens de la distance (i.e., D_{Moy} ou D_{Dyn}) considérée est générée. Le principe de décision d'identification d'un segment est la règle du k-Plus-Proche-Voisin (k=3) avec vote majoritaire. Selon ce principe, les résultats de reconnaissance se répartissent en 3 catégories : Identification, Substitution et Rejet.

Les résultats de reconnaissance (cf. Tableaux 1) sont donnés globalement pour les deux locuteurs, pour chacune des segmentations et chacune des distances testées.

DT	Ident	Subst	Rejet
Moy	71,9%	17,3%	10,8%
Dyn	80,4%	12,7%	6,8%

RM	Ident	Subst	Rejet
Moy	75,2%	16,4%	8,4%
Dyn	77,8%	14,6%	7,6%

Tableaux 1. Reconnaissance de l'ensemble des segments-référence

La comparaison dynamique donne les meilleurs résultats, au prix d'une complexité plus grande. On peut signaler que l'identification est meilleure (de plus de 5%) pour le locuteur féminin que pour le locuteur masculin. Les deux méthodes donnent des résultats comparables de l'ordre de 80%. Ces résultats encourageants nous amènent à faire le DAP sur l'ensemble des segments.

4.2. DAP de l'ensemble des segments

Dans ces expériences, pour chaque segment-test de l'ensemble des segments, l'ensemble d'apprentissage est constitué de l'ensemble des segments-référence auquel les segments-référence trop proches temporellement du segment-test ont été éliminés. Le test de proximité permet d'éviter tout effet de bord qui biaiserait les résultats ; le seuil de proximité a été fixé à 0.5 s.

Le principe d'identification du segment-test est le même que celui défini précédemment (cf. §4.1). Le processus de DAP fournit une chaîne phonétique. L'alignement de Wagner et Fisher avec la transcription phonétique normative permet de calculer les taux de DAP : Identification, Substitution, Omission et Insertion. Ces taux sont exprimés en pourcentage dans les tableaux suivants (cf. Tableaux 2).

Ces résultats confirment les performances obtenues par la comparaison dynamique. Les taux d'identification et de substitution sont très proches des résultats obtenus par l'évaluation de la cohérence des segments-référence. Les taux d'insertion sont importants (i.e., 26% pour DT, et 46% pour RM) et dégraderaient toute étape ultérieure de reconnaissance.

DT	Ident	Subst	Omi	Inser
Moy	73,1%	16,6%	10,3%	25,8%
Dyn	79,9%	12,2%	7,8%	25,9%

RM	Ident	Subst	Omi	Inser
Moy	78,2%	16,2%	5,5%	48,4%
Dyn	79,6%	15,4%	4,9%	46,1%

Tableaux 2. DAP de l'ensemble des segments

Une analyse des résultats par phonème (cf. Tableau 4) montre que les meilleurs taux d'identification sont obtenus par les deux méthodes sur les fricatives non-voisées (≈96%), les voyelles nasales (≈90%), les consonnes nasales (≈86%), les fricatives voisées (≈85%). On peut remarquer aussi que 20% des omissions proviennent des liquides, et que 40% des insertions sont observées pour [p], [t] et [v].

Aucune technique d'apprentissage n'est utilisée : chaque segment est comparé de façon exhaustive avec tous les autres. Néanmoins ce DAP permet d'évaluer les méthodes de segmentation du signal de parole sans introduire le biais d'une méthode d'apprentissage.

4.3. Conclusion

Les deux méthodes DT et RM présentent des taux d'identification, de substitution et d'omission comparables et ne permettent pas d'écarter une méthode au profit de l'autre pour le DAP. Nous cherchons par conséquent à exploiter notre hypothèse de complémentarité des deux méthodes.

5. DAP PAR INTEGRATION DES METHODES DT ET RM

A priori, les segments-référence sont les segments qui devraient être les mieux reconnus puisqu'ils contiennent par définition la réalisation d'un phonème. Les autres segments sont des réalisations intermédiaires qui peuvent donner soit une identification identique à un segment voisin, soit une identification fausse. Dans le premier cas, les segments peuvent être fusionnés, dans le second cas, l'erreur est irrémédiable. L'étiquetage manuel ne localise que le centre des réalisations phonétiques. Il ne donne aucune information sur les zones de transition, il ne permet par conséquent aucun apprentissage sur celles-ci. Ces zones transitoires sont à l'origine de nombreuses insertions, que le principe de décision sur les k-Plus-Proche-Voisins ne peut pas toujours écarter.

Nous faisons l'hypothèse suivante :
- les segments non-transitoires seront bien identifiés par les deux méthodes,
- les zones transitoires seront soit rejetées par une des méthodes, soit identifiées différemment.

5.1. Méthode d'intégration et résultats

La méthode consiste à construire une nouvelle segmentation composée de toutes les intersections entre les segments des deux méthodes. Chaque nouveau segment est étiqueté d'après les 6-Plus-Proches-Voisins (i.e., 3-PPV pour chacune des méthode), le seuil de rejet étant fixé à 4 (cf. Figure 4).

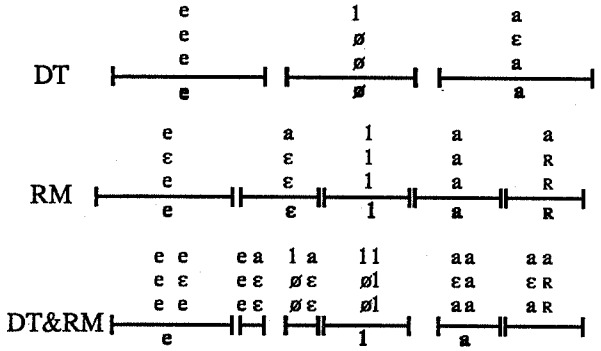


Figure 4. Intégration des deux méthodes sur le segment de parole [ela]

Les résultats de ce nouveau DAP sont donnés au tableau 3. Le tableau 4 présente les résultats par phonème des taux de reconnaissance pour les trois méthodes (i.e., DT, RM et DT&RM). La distance utilisée est la comparaison dynamique.

DT & RM	Ident	Subst	Omi	Inser
Moy	66,9%	10,7%	22,4%	8,4%
Dyn	71,7%	8,0%	20,3%	6,7%

Tableau 3. DAP par intégration des méthodes DT et RM

Ces résultats confirment notre hypothèse précédente : le taux d'insertion diminue en moyenne de 36% à 7% alors que le taux d'identification ne baisse que de 80% à 72%, le taux d'erreur étant passé en moyenne de 56% à 35%. Il y a cependant beaucoup plus d'omissions. L'étude des segments non-étiquetés devrait permettre de diminuer les taux d'omission (e.g., discrimination d'après la longueur de ces segments).

6. CONCLUSIONS

Nous avons présenté une approche coopérative entre deux méthodes. Bien que cette approche soit pour l'instant générale et ne fasse appel à aucune connaissance phonétique, notre objectif est en partie atteint. En effet, la segmentation produite par intersection des deux premières segmentations permet d'obtenir des "îlots de confiance". La plupart des insertions a été éliminée. L'utilisation de ces îlots et d'autres sources d'information (e.g., voisement, explosions) devrait permettre d'augmenter les taux d'identification grâce à des connaissances acquises par un système d'apprentissage automatique. Cela devrait renforcer la synergie des méthodes et permettre d'obtenir un DAP robuste.

Cette étude a été réalisée dans le cadre d'un projet GRECO-Reconnaissance réunissant le LAFORIA, l'ENST et l'IRISA.

REFERENCES

[1] B.S. ATAL, "Efficient Coding of LPC Parameters by Temporal Decomposition", IEEE ICASSP, pp. 81-84, 1983.
[2] R. ANDRE-OBRECHT, "A New Statistical Approach for the Automatic Segmentation of Continuous Speech Signals", IEEE Trans. ASSP, vol 36, pp. 29-40, janvier 1988.
[3] G. CHOLLET, Y. GRENIER & S.-M. MARCUS, "Temporal Decomposition and Non-Stationary Modeling of Speech", Eurasip, pp. 365-368, 1986.
[4] F. BIMBOT, G. CHOLLET, P. DELEGLISE & C. MONTACIE, "Temporal Decomposition and Acoustic-Phonetic Decoding of Speech", IEEE ICASSP, pp. 445-448, New-York, USA, 1988.
[5] P. DELEGLISE, "Une architecture logicielle pour le décodage acoustico-phonétique. Application à la détection d'événements phonétiques". Thèse d'Etat, Paris VI, France, 1991.
[6] L.J. BOE & L. MICLET, "Manuel d'étiquetage large. GRECO n°39 Communication Parlée", Commission étiquetage BDSON & EUROM, 1988.
[7] DAVIS & MERMELSTEIN, "Comparaison of Parametric Representations for Monosyllabic Word Recognition in Continuous Spoken Sentences", IEEE ASSP, vol 28, 1980.
[8] E. ZWICKER, "Masking Period Patterns and Hearing Theories in Psychophysics ans Physiology of Hearing", Ed. E.F. Evans & J.P. Wilson, New-York, 1977.
[9] C. MONTACIE, "Décodage acoustico-phonétique : apport de la décomposition temporelle généralisée et de transformaions spectrales non-linéaires". Thèse de l'ENST, 1991.

		DT				RM				DT & RM			
Phon	Nbr	Ident	Subst	Omi	Inser	Ident	Subst	Omi	Inser	Ident	Subst	Omi	Inser
p	179	63%	31%	6%	124%	54%	42%	4%	265%	35%	23%	42%	13%
t	357	81%	12%	8%	100%	85%	12%	3%	145%	74%	6%	20%	24%
k	186	70%	23%	7%	44%	80%	18%	2%	112%	61%	18%	21%	15%
b	163	71%	21%	8%	10%	63%	34%	3%	29%	53%	16%	31%	2%
d	429	85%	7%	8%	17%	86%	10%	5%	38%	79%	4%	17%	5%
g	148	45%	46%	9%	11%	39%	56%	5%	19%	25%	38%	37%	1%
m	238	85%	12%	3%	9%	75%	22%	3%	13%	69%	13%	18%	3%
n	251	86%	9%	6%	20%	82%	12%	6%	41%	75%	5%	20%	4%
f	152	95%	3%	2%	7%	98%	2%	0%	11%	95%	2%	3%	0%
s	300	97%	1%	2%	7%	96%	3%	1%	8%	95%	1%	4%	1%
ʃ	154	97%	1%	2%	5%	99%	1%	0%	2%	95%	1%	3%	1%
v	221	86%	7%	7%	113%	85%	9%	5%	200%	78%	4%	18%	48%
z	165	83%	5%	12%	27%	90%	7%	3%	97%	80%	4%	16%	10%
ʒ	146	94%	3%	3%	13%	95%	5%	0%	22%	92%	2%	6%	3%
l	635	73%	11%	16%	10%	61%	20%	18%	17%	56%	8%	37%	2%
ʀ	492	62%	21%	17%	32%	79%	15%	6%	128%	57%	8%	35%	18%
a	681	88%	7%	4%	10%	87%	10%	3%	13%	83%	4%	13%	1%
i	329	85%	7%	8%	20%	87%	8%	5%	23%	81%	5%	14%	4%
e	406	84%	11%	4%	16%	83%	13%	4%	16%	74%	6%	19%	4%
ɛ	335	68%	22%	10%	21%	73%	24%	3%	15%	61%	18%	21%	3%
y	262	70%	21%	9%	16%	70%	22%	8%	9%	61%	10%	29%	2%
ø	664	81%	9%	9%	27%	83%	13%	4%	47%	77%	5%	18%	8%
u	218	75%	17%	9%	23%	68%	25%	7%	14%	62%	11%	27%	3%
o	150	67%	23%	10%	19%	65%	26%	9%	11%	57%	16%	27%	4%
ɔ	221	70%	24%	6%	25%	66%	27%	6%	14%	61%	18%	21%	3%
œ̃	294	93%	2%	4%	16%	88%	8%	4%	13%	87%	2%	10%	3%
ɛ̃	229	86%	8%	6%	10%	86%	12%	2%	14%	78%	7%	15%	2%
ɔ̃	215	90%	5%	5%	12%	93%	6%	2%	10%	88%	4%	8%	3%
Total	8220	79,9%	12,2%	7,8%	25,9%	79,6%	15,4%	4,9%	46,1%	71,7%	8,0%	20,3%	6,7%

Tableau 4. Résultats analytiques des taux de reconnaissance pour les trois méthodes (DT, RM et DT&RM).

COMMANDE D'UN MODELE DU CONDUIT VOCAL BASÉE SUR L'ESTIMATION DES FONCTIONS DE SENSIBILITÉ

Paul JOSPA*, Alain SOQUET* & Marco SAERENS**

*Université Libre de Bruxelles
**Lernout & Hauspie Speech Products et Université Libre de Bruxelles

RESUME:
Les réseaux connexionistes sont actuellement utilisés dans le cadre de l'inversion acoustico-articulatoire. Ainsi, nous avons récemment utilisé la théorie des Régions et des Modes Distinctifs (Mrayati, Carré & Guérin, 1988), en combinaison avec un réseau connexionniste pour réaliser une inversion acoustico-articulatoire d'un modèle du conduit vocal. L'algorithme utilisé, une variante de l'algorithme de rétro-propagation de l'erreur, a été développé dans le cadre de la commande de processus, et présente de fortes similarités avec les algorithmes de commande adaptative. Il nécessite la connaissance d'une approximation du Jacobien du processus physique. Dans cet article, nous présentons des comparaisons entre les configurations obtenues à partir des deux approximations différentes (Jacobien constant, évalué à partir de la théorie des Régions Distinctives, et Jacobien variable obtenu à partir d'une approximation des fonctions de sensibilités). Nous comparons les résultats avec les configurations proposées par (Majid, Boe & Perrier, 1986) pour le Français,et celles publiées par (Fant, 1960) pour le Russe.

INTRODUCTION

Les réseaux connexionistes sont actuellement utilisés dans le cadre de l'inversion acoustico-articulatoire (Bailly, Bach, Laboissiere & Olesen, 1990; Laboissière, Schwartz & Bailly, 1991; Shirai & Kobayashi, 1991; Rahim, Kleijn, Schroeter & Goodyear, 1991). Ainsi, nous avons récemment utilisé la théorie des Régions Distinctives et des Modes (Mrayati, Carré & Guérin, 1988), en combinaison avec un réseau de neurones artificiels (Saerens & Soquet, 1991) pour réaliser une inversion acoustico-articulatoire d'un modèle du conduit vocal (Soquet, Saerens & Jospa, 1990, 1991). L'algorithme utilisé, une variante de l'algorithme de rétro-propagation de l'erreur, a été développé dans le cadre de la commande de processus (Saerens, Soquet, Renders & Bersini, 1992; Saerens, 1991), et présente de fortes similarités avec des algorithmes de commande adaptative. Il nécessite la connaissance à priori d'une approximation du Jacobien du processus physique. Lors de nos precédentes expériences, l'approximation du Jacobien était constante et explicitement fournie par la théorie des régions distinctives et des modes. Les configurations obtenues étaient satisfaisantes pour la plupart des voyelles orales du Français --nous utilisons les valeurs publiées par (Majid, Boe & Perrier, 1986)--, à l'exeption du [u]. La raison en est que le modèle du conduit utilisé doit être modifié pour produire cette voyelle de manière adéquate (Carré & Mrayati, 1991). L'utilisation d'un modèle spécifique au [u] ne peut cependant pas s'intégrer naturellement dans l'algorithme d'inversion que nous utilisons. Dès lors, nous nous sommes proposés de calculer en ligne une approximation du Jacobien correspondant à l'état actuel du conduit vocal. Ceci peut se faire en perturbant successivement les différents paramètres articulatoires du conduit vocal et en calculant les modifications correspondantes des valeurs formantiques. Nous testons cette méthode sur les voyelles orales du Français (Majid, Boe & Perrier, 1986) et les voyelles du Russe (Fant, 1960). Les configurations obtenues sont plus satisfaisantes. En particulier, les fonctions d'aire obtenues pour la voyelle [u] sont beaucoup plus proches des configurations de référence publiées. Des méthodes moins coûteuses en temps de calcul sont en cours d'évaluation.

Après avoir brièvement rappelé la méthode utilisée pour l'inversion, nous présentons des comparaisons entre les configurations obtenues selon que le Jacobien est évalué à partir de la théorie des Régions Distinctives, ou à partir d'une estimation des fonctions de sensibilité. Les configurations proposées par (Majid, Boe & Perrier, 1986) pour le Français, et celles publiées par (Fant, 1960), pour le Russe, servent de réference.

ALGORITHME D'APPRENTISSAGE

La méthode d'inversion utilisée a été proposée dans le cadre plus général de la commande de processus (*Figure 1*). Cette architecture est motivée par les capacités d'approximation universelle des réseaux multi-couches (Hornik, Stinchcombe & White, 1989), et par le fait

qu'un tel réseau, entraîné à l'aide d'un critère adéquat, approche l'espérance conditionnelle $E(y|x)$ d'observer la sortie y, étant donné l'entrée x (White, 1989).

Dans notre cas (*Figure 1*), la dernière couche du réseau fournit à chaque instant k les entrées $u_\alpha(k)$ du processus (les paramètres de commande), lequel donne ensuite, en sortie: $y_\beta(k+1)$, et non les cibles désirées: $y^d_\beta(k+1)$. Dans le cas d'un processus d'ordre p, en plus des sorties désirées $y^d_\beta(k+1)$, le reseau reçoit en entrée les p derniers vecteurs de sortie du processus, ainsi que les p derniers paramètres de commande, de manière à pouvoir reconstruire l'état du processus. Le réseau doit donc fournir successivement les paramètres de commande $u_\beta(k-1)$, $u_\beta(k-2)$, ..., $u_\beta(k-p)$ qui minimisent : $E(k) = \frac{1}{2}\sum_\alpha (y_\alpha(k) - y^d_\alpha(k))^2$. L'optimisation s'effectue habituellement à l'aide d'une descente de gradient dans l'espace $E(w_{\alpha\beta})$ (Le Cun, 1985; Rumelhart, Hinton & Williams, 1986). Afin de calculer le gradient, nous introduisons la fonction de Lagrange $L(y_\alpha(k), u_\alpha(k-i), U^l_\alpha(k-i), w^l_{\alpha\beta}(k-i))$ définie à chaque instant k :

$$L(k) = \frac{1}{2}\sum_\alpha (y_\alpha(k) - y^d_\alpha(k))^2 \quad (1a)$$

$$+ \sum_\alpha \Gamma_\alpha(k)\ (y_\alpha(k) - g_\alpha[y_\beta(k-1), ..., y_\beta(k-p), u_\beta(k-1), ..., u_\beta(k-p)]) \quad (1b)$$

$$+ \sum_{i=1}^{p} \sum_\alpha \Lambda_\alpha(k-i)\ (u_\alpha(k-i) - U^q_\alpha(k-i)) \quad (1c)$$

$$+ \sum_{i=1}^{p} \sum_{l,\alpha} \lambda^l_\alpha(k-i)\ (U^l_\alpha(k-i) - f[\sum_\beta w^l_{\alpha\beta}(k-i)\ .\ U^{l-1}_\beta(k-i)]) \quad (1d)$$

où $U^l_\alpha(k-i)$ est l'activation de l'unité α à la couche l et l'instant $(k-i)$ et $w^l_{\alpha\beta}(k-i)$ est le poids de la connexion reliant l'unité α de la couche l à l'unité β de la couche $l-1$, également à l'instant $(k-i)$. Le premier terme de l'équation (1a) correspond à la fonction coût, le deuxième (1b) à la représentation entrée-sortie d'un processus d'ordre p, le troisième (1c) impose simplement que les unités de la dernière couche $U^q_\alpha(k)$ (dernière couche q) fournissent les paramètres de commande $u_\alpha(k)$ à chaque instant k, et le dernier (1d) correspond à la fonction de transfert des unités. $\Gamma_\gamma(k)$, $\Lambda_\gamma(j)$, $\lambda^l_\gamma(j)$ sont des multiplicateurs de Lagrange. Les conditions d'extrémum : $\frac{\partial L}{\partial y_\gamma(k)} = 0$, $\frac{\partial L}{\partial u_\gamma(j)} = 0$, $\frac{\partial L}{\partial U^l_\gamma(j)} = 0$, nous permettent de calculer les valeurs des multiplicateurs de Lagrange $\Gamma_\gamma(k)$, $\Lambda_\gamma(j)$, $\lambda^l_\gamma(j)$ (Le Cun, 1989). En revanche, les valeurs des poids des connexions s'obtiennent par descente de gradient : $w_{\alpha\beta}(j+1) = w_{\alpha\beta}(j) - \eta \frac{\partial L(k)}{\partial w_{\alpha\beta}(j)}$. De cette manière, nous obtenons pour l'erreur rétro-propagée en dernière couche:

$$\delta^q_\gamma(j) = f'^q_\gamma(j) \sum_{k=j+1}^{j+p} \sum_\alpha (y_\alpha(k) - y^d_\alpha(k)) \frac{\partial g_\alpha(k)}{\partial u_\gamma(j)} \quad (2)$$

(pour plus de détails, voir Saerens, 1991; une autre dérivation peut être trouvée dans Saerens & Soquet, 1991). Nous pouvons remarquer que le calcul du gradient fait appel au Jacobien du processus $\frac{\partial g_\alpha(k)}{\partial u_\gamma(j)}$, qui est inconnu (dans le cas de la méthode proposée par Jordan, le Jacobien est évalué par rétro-propagation à travers un réseau identifiant le processus : Jordan, 1989).

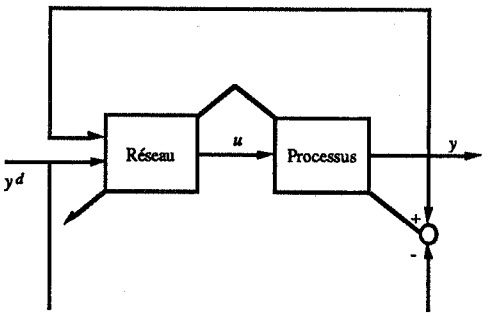


Figure 1 : Architecture d'apprentissage en boucle fermée.

Pour un processus du premier ordre, et si les termes du Jacobien ne changent pas de signe, nous pouvons simplement utiliser la règle :

$$\delta^q_\gamma(k) = f'^q_\gamma(k) \sum_\alpha \xi_\alpha\, sgn(\frac{\partial y_\alpha(k+1)}{\partial u_\gamma(k)})\ .\ (y_\alpha(k+1) - y^d_\alpha(k+1)) \quad (3)$$

où ξ_α est une variable stochastique de moyenne *1.0*, uniformément distribuée sur l'intervalle *[0.5, 1.5]* (Saerens & Soquet, 1991; Saerens, 1991). Cette

approximation diminue l'erreur car le produit scalaire entre le gradient réel et l'approximation (3) est positif. Cet algorithme sera utilisé pour l'entraînement de notre réseau.

MÉTHODE D'INVERSION

Nous avons utilisé l'algorithme défini précédemment afin de réaliser l'inversion du modèle de production proposé par (Mrayati, Carré & Guérin, 1988), qui se base sur un tube acoustique à *8* régions de longueurs inégales (*Figure 2*).

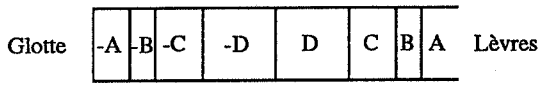


Figure 2 : Tube acoustique subdivisé en 8 régions de longueurs inégales.

L'entraînement est réalisé sur l'entièreté de l'espace vocalique. En adjoignant la contrainte d'un volume constant pour le tube acoustique, la convergence ne pose pas de problème (pour plus de détails, voir Soquet, Saerens & Jospa, 1990, 1991 ou Saerens, 1991). La première section (-A) est, par ailleurs, limitée à l'intervalle [*1.0 cm*^{2}, *2.5 cm*^{2}]. La fonction optimisée s'écrit:

$$E = \frac{1}{2} \sum_{v} \sum_{i=1}^{3} (F_i^v - F_i^{vd})^2 + k \sum_{v} [(\sum_{i=1}^{8} L_i A_i^v) - V_0]^2 \quad (4)$$

où les F_i^v représentent les formants associés au tube; les F_i^{vd} les formants cibles; les L_i les longueurs des différentes régions du tube; les A_i^v les sections associées qui sont fournies par le réseau; et V_0 le volume total du tube. La somme porte sur l'ensemble des voyelles v générées dans l'espace vocalique.

Le processus d'apprentissage se déroule comme suit (*Figure 3*):

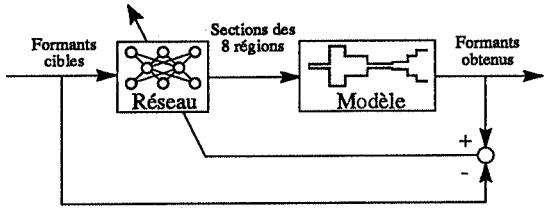


Figure 3 : Le réseau doit fournir les sections du tube qui permettent de produire les formants cibles.

1. Le réseau reçoit en entrée les formants cibles F_i^{vd}, générés dans l'espace vocalique.
2. Les sorties du réseau fournissent les aires des sections du tube A_i^v, et les valeurs correspondantes des formants F_i^v sont calculées (pour ce calcul, nous utilisons un algorithme développé par Liljencrants & Fant, 1975).
3. Les différences entre ces valeurs et les formants cibles sont utilisées pour calculer l'erreur en sortie, et corriger les poids en rétro-propageant par (2) ou (3). Retour en 1.

Ainsi, après entraînement, le réseau approche la fonction reliant les valeurs formantiques désirées, en entrée, et les configurations du tube, en sortie.

RÉSULTATS DES SIMULATIONS

Nous présentons en *Figure 4* les configurations obtenues pour certaines voyelles orales du Français (valeurs des trois premiers formants tirées de Majid, Boe & Perrier, 1986). La *première configuration*, représentée à gauche, correspond à la configuration proposée par (Majid, Boe & Perrier, 1986). La *deuxième configuration* représentée (signs) a été obtenue en utilisant les signes des dérivées partielles du Jacobien, qui ne sont, en principe, valables qu'aux alentours du tube neutre; nous avons donc utilisé (3) afin de minimiser (4). Les signes sont tirés de l'étude réalisée par (Mrayati, Carré & Guérin, 1988). Ces formes du conduits sont donc celles obtenues en sortie du réseau, après convergence sur l'entièreté de l'espace vocalique. Le tube acoustique a une longueur constante de *19 cm.*

Les *configurations 3 et 4* (à droite) ont été obtenues en évaluant, à chaque itération, le Jacobien, à l'aide de petites perturbations de la fonction d'aire localisées aux régions (fonctions de sensibilité). Cette méthode est évidemment fort coûteuse en temps de calcul, et nous envisageons d'utiliser des algorithmes plus économiques (voir conclusion). La troisième configuration correspond à un tube de *19 cm* et la quatrième à un tube de *17 cm.* Les différentes formes sont qualitativement fort semblables, mais, comme prévu, la voyelle [u] est mieux modélisée avec le calcul des fonctions de sensibilité.

En *Figure 5*, nous présentons les configurations obtenues pour les voyelles du Russe (Fant, 1960). La configuration la plus à gauche, est celle publiée par (Fant, 1960). Ici, seuls les tubes acoustiques obtenus via les sensibilités locales sont représentés, et ce pour trois longueurs: *19 cm, 18 cm* et *17 cm.*

CONCLUSION

Les formes du conduit obtenues sont qualitativement acceptables. Cependant, l'estimation des fonctions de sensibilité par perturbations des régions demande un temps de calcul très important. Nous envisageons deux solutions à ce problème. La première consiste à utiliser la méthode variationnelle de Jospa (1972, 1991). En plus des formants, cette méthode fournit explicitement, et avec trés peu de calculs supplémentaires, les fonctions de sensibilité associées à la configuration proposée par le réseau. La seconde solution consisterait à exploiter des relations implicites (Bonder, 1983 ou Jospa, 1972, 1991) reliant valeurs des formants et sections du tube. Bien qu'il soit

difficile de résoudre ces relations non-linéaires, celles-ci peuvent néanmoins être avantageusement utilisées pour évaluer le gradient de l'erreur. En effet, nous pouvons introduire ces relations en lieu et place du terme (1b) du Lagrangien, et recalculer le gradient.

RÉFÉRENCES

BAILLY G., BACH M., LABOISSIERE R. & OLESEN M., 1990, "Generation of articulatory trajectories using sequential networks". Proceedings of the First ESCA Workshop on Speech Synthesis, Autrans, pp. 67-70.

BONDER L., 1983, "The n-tube formula and some of its consequences". Acustica, 52, pp. 216-226.

CARRÉ R. & MRAYATI M., 1991, "Static and dynamic relations between vocal tract configurations and acoustics". Proceedings of the XIIth International Congress of Phonetic Sciences, Aix-en-Provence, pp. 210-214.

FANT G., 1960, "Acoustic theory of speech production". Mouton & Co.

HORNIK K., STINCHOMBE M. & WHITE H., 1989, "Multilayer feedforward networks are universal approximators". Neural Networks, 2, pp. 359-366.

JORDAN M.I. , 1989, "Generic constraints on underspecified target trajectories". Proceedings of International Joint Conference on Neural Networks, Washington, Vol I, pp. 217-225.

JOSPA P., 1972, "Forme approchée du conduit vocal déduite des fréquences de résonnance. Théorie des perturbations et méthode variationnelle". Actes des 5èmes J.E.P., Lannion 1972. pp. 225-261.

JOSPA P., 1991, "Des paramètres formantiques au profil articulatoire". Proceedings of the XIIth International Congress of Phonetic Sciences, Aix-en-Provence, pp. 378-381.

LABOISSIERE R., SCHWARTZ J.-L. & BAILLY G., 1991, "Motor control for speech skills: A connectionist approach". Proceedings of the 1990 Connectionnist Models Summer School, Touretzky D., Elman J., Sejnowski T. & Hinton G. (editors), Morgan Kaufmann Publishers, pp. 319-327.

LE CUN Y., 1985, "A learning scheme for asymmetric treshold networks". Proceedings of Cognitiva 85, Paris, pp. 599-604.

LE CUN Y., 1989, "A theorical framework for back-propagation". Proceedings of the 1988 Connectionnist Models Summer School, Touretzky D., Hinton G. & Sejnowski T. (editors), Morgan Kaufmann Publishers, pp. 21-28.

LILJENCRANTS J. & FANT G., 1975, "Computer program for VT-resonance frequency calculations". Stockholm : Speech Transmission Laboratory – Quarterly Progress and Status Report, 4/1975, pp. 15-20.

MAJID R., BOE L.J. & PERRIER P., 1986, "Fonctions de sensibilité, modèle articulatoire et voyelles du Français". Actes des 15èmes Journées d'Etude sur la Parole, Aix en Provence, pp. 59-63.

MRAYATI M., CARRÉ R. & GUÉRIN B., 1988, "Distinctive regions and modes: A new theory of speech production". Speech Communication, 7, pp. 257-286.

RAHIM M.G., KLEIJN W.B., SCHROETER J. & GOODYEAR C.C., 1991, "Acoustic to articulatory parameter mapping using an assembly of neural networks". Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Toronto, pp. 485-488.

RUMELHART D.E., HINTON G.E. & WILLIAMS R.J., 1986, "Learning internal representation by error propagation". In Distributed Parallel Processing, Vol. 1, MIT Press, Cambridge, pp. 318-362.

SAERENS M., 1991, "Approche connexionniste de la commande de processus. Application à l'inversion acoustico-articulatoire d'un modèle de conduit vocal, en vue de l'identification du lieu d'articulation des occlusives du Français". Thèse de Doctorat en Sciences Appliquées, Université Libre de Bruxelles, Institut de Phonétique.

SAERENS M. & SOQUET A., 1991, "Neural Controller Based on Back-Propagation Algorithm". IEE Proceedings–F, 138(1), pp. 55-62.

SAERENS M., SOQUET A., RENDERS J-M. & BERSINI H., 1992, "A preliminary comparison between a neural adaptive controller and a model reference adaptive controller". To appear in the Proceedings of the International Workshop on Neural Networks in Robotics; Kluwer.

SHIRAI K. & KOBAYASHI T., 1991, "Estimation of articulatory motion using neural networks". Journal of Phonetics, 19, pp. 379-385.

SOQUET A., SAERENS M. & JOSPA P., 1990, "Acoustic-articulatory inversion based on a neural controller of a vocal tract model". Proceedings of the ESCA First International Workshop on Speech Synthesis, Autrans, pp. 71-74.

SOQUET A., SAERENS M. & JOSPA P., 1991, "Acoustic-articulatory inversion based on a neural controller of a vocal tract model: Further results". Proceedings of the International Conference on Artificial Neural Networks, Helsinki; Kohonen, Mäkisara, Simula & Kangas (editors), North-Holland, pp. 371-376.

WHITE H., 1989, "Learning in artificial neural networks: A statistical perspective". Neural Computation, 1, pp. 425-464.

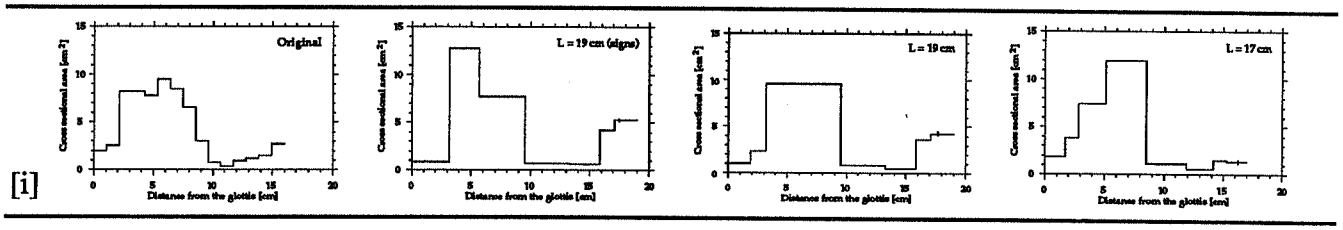
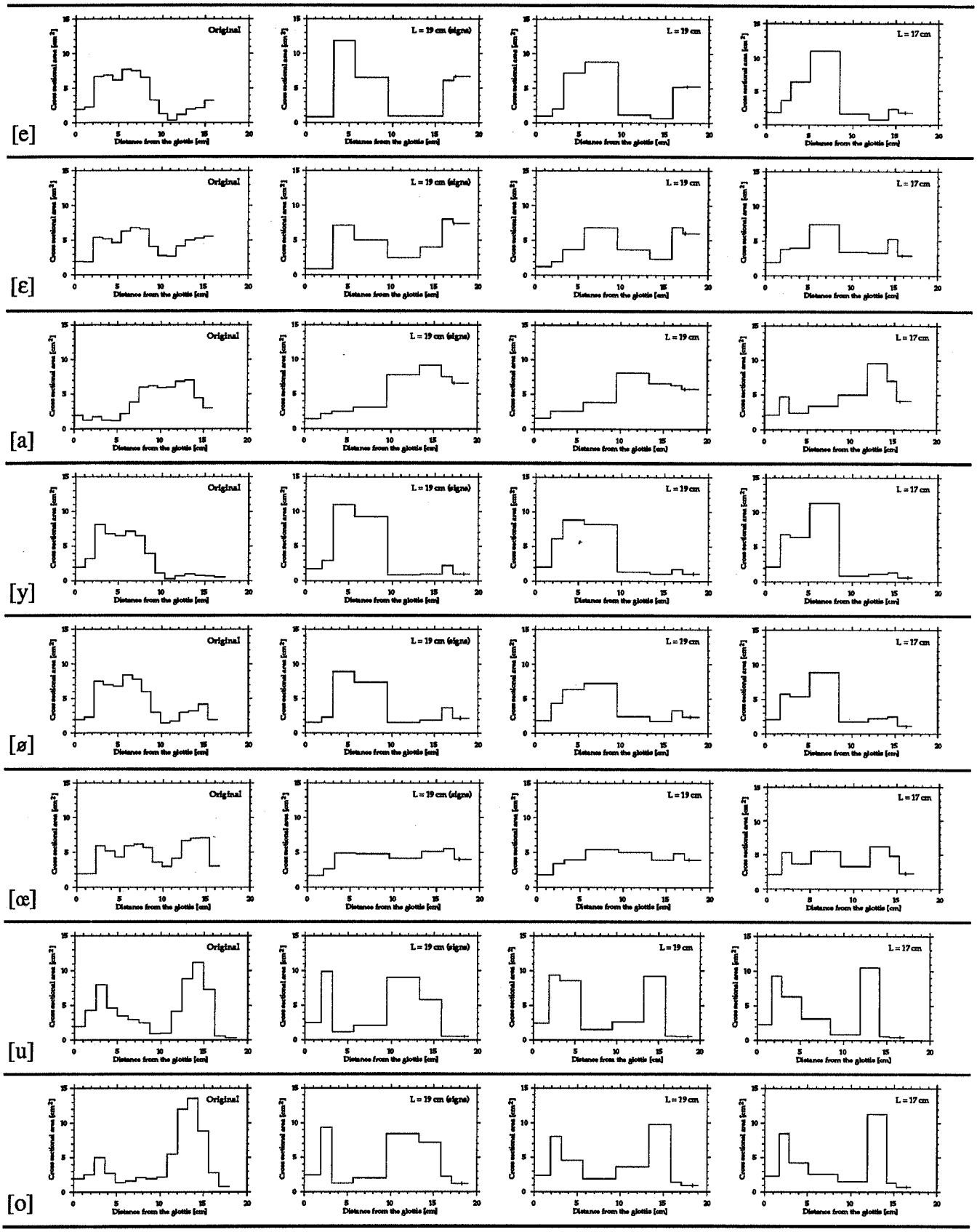


Figure 4 (voir page suivante).


[e]
[ɛ]
[a]
[y]
[ø]
[œ]
[u]
[o]
Original
L = 19 cm (elgra)
L = 19 cm
L = 17 cm
Cross sectional area [cm²]
Distance from the glottis [cm]

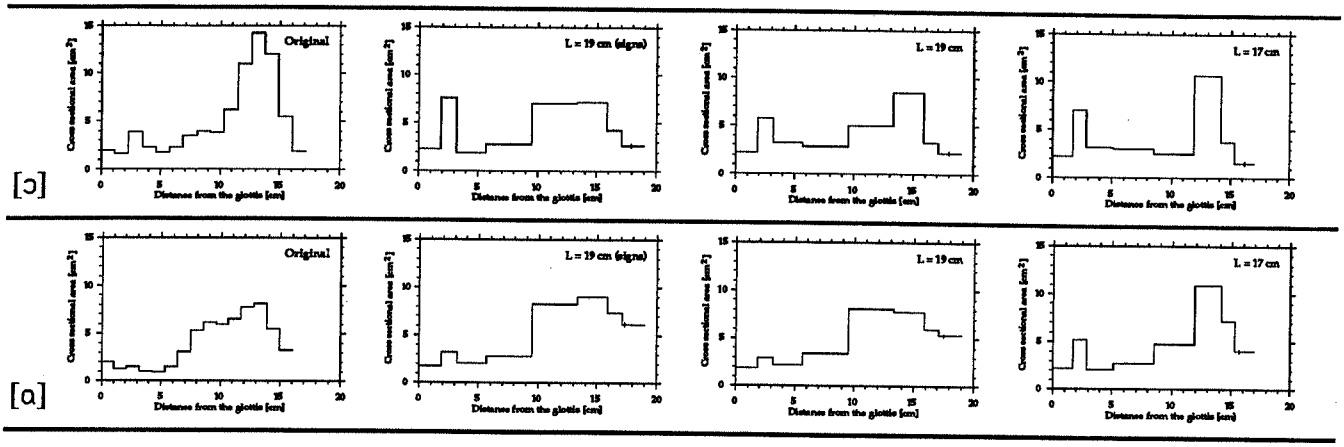


Figure 4 (suite): Configurations de tubes acoustiques. De gauche à droite: la première correspond à la référence publiée par Majid, Boe & Perrier; la deuxième est celle obtenue par le réseau en utilisant simplement le signe du Jacobien; les troisième et quatrième sont obtenues par le réseau en évaluant à chaque itération les fonctions de sensibilité.

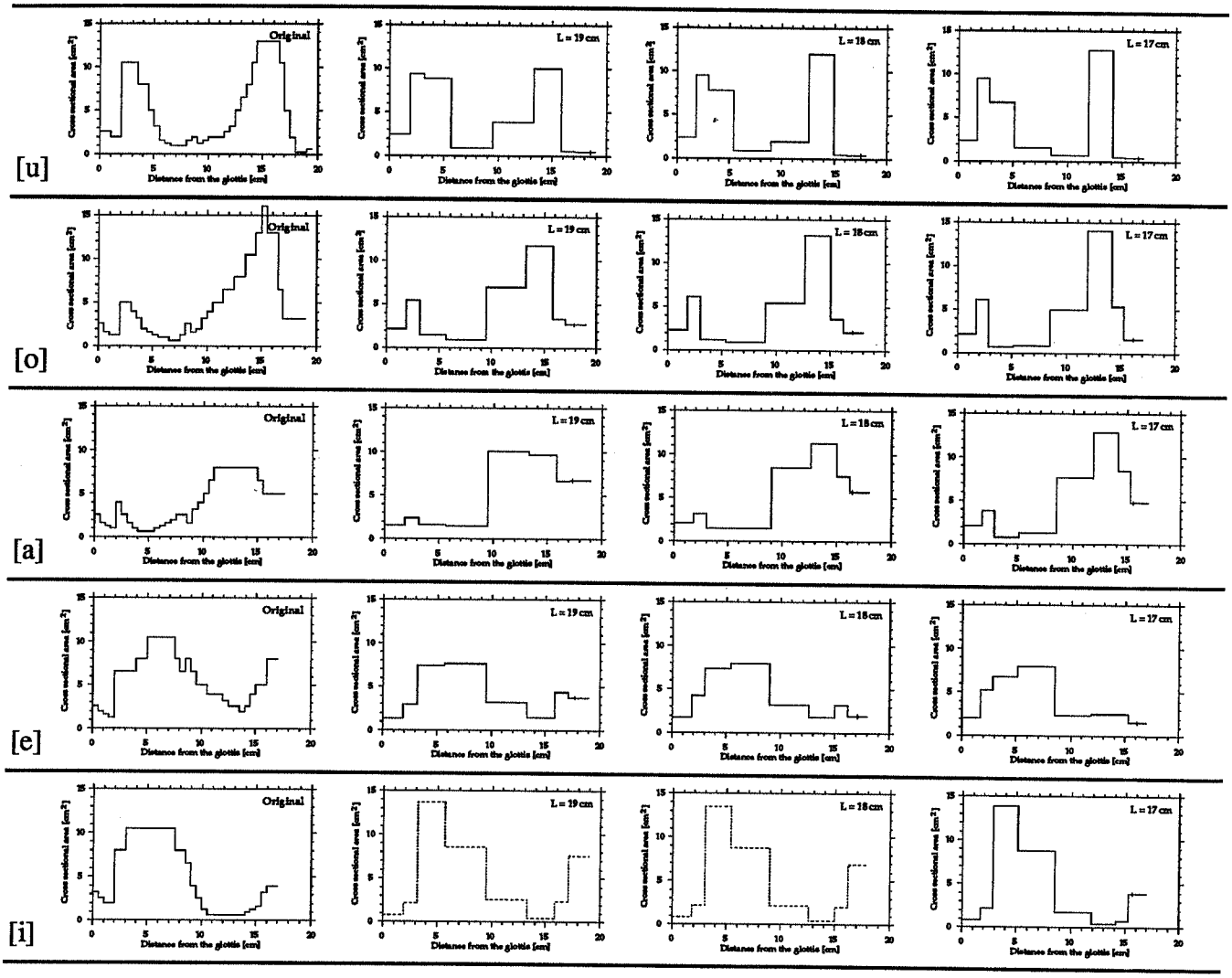


Figure 5 : Configurations de tubes acoustiques; la plus à gauche correspond à la référence publiée par Fant; les trois autres sont obtenues par le réseau en évaluant à chaque itération les fonctions de sensibilité.

19es J.E.P. - Bruxelles, 19 au 22 mai 1992.

LES PROBLÈMES DE PHONÉTISATION DU "e" EN CONTEXTE CONSONANTIQUE POUR UN LEXIQUE DE 59 000 MOTS

J. VAN EIBERGEN

ICP UA CNRS n° 368 INPG/ENSERG UNIVERSITÉ STENDHAL
BP 25 38 040 GRENOBLE CEDEX 9 FRANCE

Résumé

L'objectif de cette communication est de faire le point sur les problèmes que pose la phonétisation de la lettre "e", sans diacritiques et en contexte consonantique, à partir d'un dictionnaire de 59 000 mots (Le Petit Robert 1990). Ce travail est effectué dans le cadre plus général de la phonétisation des dictionnaires dont l'objectif est de proposer une description linguistique du passage de la forme orthographique vers la forme phonétique, organisé en un processus automatisable. Pour un même type d'écriture, le nombre de règles concernant la classe du "e" est supérieur à celles des autres classes alphabétiques. Ceci s'explique par la fréquence graphique de cette lettre et par une grande variation de contextes de mot régissant des réalisations phonétiques nombreuses.

La phonétisation du "e" pose le problème plus général de la transcription des variations phonétiques vocaliques mais aussi consonantiques et celui du choix d'une norme.

1. INTRODUCTION

Dans une grammaire de phonétisation du français, la notation des sons qui pose le plus de problèmes concerne en priorité les faits tels que le [a, ɑ], le [œ, ɛ̃], les géminées, l'aperture des voyelles intermédiaires, et le [ə] appelé traditionnellement caduc, muet ou latent. La lettre "e" renvoie aux deux derniers phénomènes. Pour un même type d'écriture (le langage TOPH), le nombre de règles concernant la classe de cette lettre est supérieur à celles des autres classes alphabétiques : sur 1 260 règles, 260 d'entre elles ont été élaborées, soit 20% de l'ensemble des règles du dictionnaire [2]. Ceci s'explique par la fréquence graphique de cette lettre : 35 000 mots environ des 59 000 du Petit Robert, contiennent au moins ce caractère. Le "e" se rencontre aussi dans une une grande variété de contextes orthographiques. Nous allons faire le point sur les problèmes que pose la phonétisation de la lettre "e" sans diacritiques et en contexte consonantique, pour le lexique de 59 000 mots du Petit Robert 1990 (PR 90) qui constitue une référence orthographique et qui propose aussi une forme phonétiqué [10]. Cette étude est effectuée dans le cadre plus général de la phonétisation des dictionnaires dont l'objectif est de proposer une description linguistique du passage de la forme orthographique vers la forme phonétique, organisé en un processus automatisable [1].

2. CHOIX MÉTHODOLOGIQUES

Dans l'écriture des règles de phonétisation du "e", nous avons choisi de respecter la notation phonétique proposée par le P.R 90. dont les choix fondamentaux de transcription phonétique évoluent et respectent davantage le français "tel qu'on le parle".

Mais l'établissement de règles fait apparaître des "incohérences" pour certaines formes phonétiques proposées par le PR 90. Il est possible de proposer des "régularisations" à ces "incohérences" avec le double objectif, difficilement conciliable, de transcrire le plus de variations phonétiques possibles et de limiter au minimum le nombre de règles pour un même type de contexte et ceci, pour des raisons d'économie.

Par ailleurs, pour nous aider encore dans nos choix phonétiques, nous avons fait des contrôles, particulièrement pour les [ə] en syllabes contiguës, en demandant à plusieurs locuteurs de prononcer les mots "à problème", dans la situation de lecture de mot isolé, situation qui est celle de toute phonétisation lexicale.

Précisons enfin que, pour la lettre "e", les "incohérences" de transcription relevées dans le PR 90 concernent principalement les variations d'aperture [e, ɛ] et celles des réalisations [±] du [ə].

3. CONTEXTE GRAPHIQUE du "e"

C'est le contexte consonantique droit qui détermine le timbre vocalique du "e" : voyelle nasale [ɑ̃]/[ɛ̃] ou orale [e/ɛ], [ə], [i], [œ], [a].

3.1. Contexte droit "m" ou "n"

A l'intérieur de mot, les suites de deux lettres *en / em* sont des digrammes [3] si elles se trouvent graphiquement suivies d'une consonne non nasale, *envie, emporter.* Lorsqu'elles sont suivies d'une consonne nasale ou d'une voyelle, elles se dissocient en deux graphèmes distincts, *flemme, venir.*

Les exceptions aux prédictions des réalisations du "e" par rapport à son contexte droit, correspondent à un changement de timbre nasal [ɑ̃ -> ɛ̃] *appendice*, ou à une dénasalisation [ɑ̃ -> ɛn/ɛm] *kendo, tempo*, et inversement [ɛn -> ɑ̃n/ɑ̃] en syllabe initiale de mot, *enivrer, ennui*. Notons, en outre, un changement de timbre vocalique exceptionnel dans les mots isolés suivants : [ɛ -> a] *femme, solennel.*

Ces exceptions se rencontrent très fréquemment dans des mots reconnaissables à leur terminaison "non française", *referendum, menchevik, kentia, alpenstock.* Cette remarque est valable pour toutes les réalisations phonétiques du "e" perçues comme non régulières.

En syllabe finale, il n'y a que le mot outil *en* qui reste un digramme. Dans tous les autres mots, il se dissocie, *spécimen, totem...*

3.2. Contexte droit : deux graphèmes consonantiques

En position interne de mot la lettre "e", sans accent, se traduit par [e/ɛ] devant le ou les graphèmes (Graph) consonantiques suivants [5] :

Graph = lettre double + {Liquide, h}, *territoire, flemmard, effluve, ecchymose*, avec { } = facultatif.

Graph = *cqu, ck , sc : becquet, meckel, descendre*
(mais pas le graphème *ch , entrechat*).

En syllabe fermée :

Graph.1 + Graph.2 + {Graph.3}, *rester, exciter, extraverti, pechblende.*

X, sexe, sexologue.

Les exceptions se rencontrent à la frontière de morphème (préfixe) et de son lexème lié ou conjoint [8], *dessus, dessous, ressasser, rescindable, restructurer.*

En syllabe finale ouverte devant une consonne muette, les contextes sont bien connus et faciles à déterminer, *er, et, ed, ef, ect, et, ets.*

En syllabe fermée de nombreux graphèmes consonantiques sont susceptibles de suivre le "e" comme dans *ver, sec, chef, net, dolmen, correct, erg, ouest...*

3.3. Contexte droit : un graphème

La lettre "e" peut se réaliser [ə] et se caractérise par une variation de réalisation phonétique [±]. A cause de cette alternance vocalique unique en français, nous qualifierons ce [ə] de bifide pour donner l'image de deux unités séparées dont chacune d'elles appartient à la même entité.

Le contexte graphique du [ə] bifide correspond à la lettre "e" sans diacritiques :

• Précédée de n graphèmes, *cheval, brebis.* Notons qu'il n'existe aucune lettre "e" à l'initiale de mot correspondant au [ə].

• Suivie — soit de {s} + #, *pote, table, arrhes.*
— soit d'un graphème + (Liquide) + (Voyelle), *tenir, entrechat, péquenot, chevreau...*

Les exceptions se situent dans les mots étrangers, *beluga, chamerops, hovercraft, sedum, seghia*, francisés, *racketteur* et dans quelques mots rares tels que, *assener, guillemeter.*

4. APERTURE DU [e/ɛ]

La variation d'aperture se réalise véritablement en syllabe ouverte ou fermée accentuée, soit en syllabe finale de mot, lorsque celle-ci porte l'accent. En syllabe ouverte non finale de mot, les oppositions d'aperture tendent à se neutraliser [14]. En français parisien, la voyelle a même tendance à la fermeture. Cette tendance à la fermeture est observée aussi dans un bon nombre de nos "parlers français" et n'apparaît toujours pas dans le PR 90. On y constate, en outre, que pour un même contexte consonantique droit, la transcription phonétique est variable.

4.1. Le "e" suivi de consonnes doubles en syllabe ouverte

Dans le PR 90, le "e" suivi d'une double consonne en syllabe interne de mot est transcrit [ɛ], [e/ɛ] ou [e] même après le graphème correspondant à la double lettre "rr" :

[ɛ] *terrible, terrer,*
[e/ɛ] *pellagre, tellurique, perruche,*
[e] *serrer, tellière, flemmard.*

On peut vraiment parler d'incohérence dans le PR90 quand le "e" a été transcrit différemment dans un même contexte consonantique, un même entourage vocalique et pour deux mots contenant le même lexème :

[e] dans *terrasser*,
[ɛ] dans *terrassier.*

Trois raisons principales expliquent la variabilité d'aperture de cette voyelle dans ce type de contexte :

• Elle se trouve en syllabe ouverte interne de mot et devrait donc être fermée.

• Elle a tendance à se fermer si, dans la syllabe suivante la voyelle est antérieure et fermée [i, e, y]. Il s'agit là d'un phénomène combinatoire d'harmonisation vocalique [7].

• Elle peut aussi se réaliser ouverte par association de famille de mot. Le "e" de *terrasse* peut se réaliser [ɛ] comme dans *terre.*

Nous opterons pour la transcription d'un son [E] dont le degré d'aperture ne serait pas fixe et intermédiaire mais au contraire variable entre [e] et [ɛ] que nous appellerons variant.

4.2. Le "e" en finale de mot devant une consonne non réalisée

Il s'agit des "e" dans des mots comme *et, mets, manger...* On constate une tendance à la fermeture même dans les mots entrant en opposition phonologique. En d'autres termes, il n'y aurait de moins en moins de différence entre *ces* et *c'est* dans des parlers français de plus en plus nombreux. Pour rendre compte de cette évolution du français, voici ce que nous proposons :
[e] le "e" dans ***e t, boulanger, chez, pied, clef, ces,***
[E] variant dans *cadet, mets, aspect.*

4.3. Mots d'emprunt

Dans ces mots "étrangers" ou d'origine étrangère, la notation phonétique du PR 90 suit les règles de transcription des mots français quant à l'aperture du [E] en syllabe ouverte interne, soit [e] majoritairement dans *beluga, select, plenum, pietà, steward.*

Nous suivrons cette notation en faisant remarquer qu'une des tendances de notre comportement linguistique est de vouloir "franciser" rapidement les mots d'emprunt par la normalisation de leur transcription dans un dictionnaire. Il n'est pas sûr que, dans les faits phonétiques, la transcription de l'aperture de cette voyelle soit toujours aussi constante. C'est ce que nous avons noté avec les contrôles auxquels nous avons procédé.

5. RÉALISATIONS [±] DU [ə]

La réalisation variable du [ə] dépend de nombreux facteurs tels que :
• sa place dans le mot,
• le nombre de ses réalisations dans le même mot en syllabes contiguës,
• le nombre et le type de consonnes le précédant et le suivant,
• le type de lexique appartenant à des niveaux de langue différents ou à un français de spécialité (français courant, littéraire, technique...)
• la situation de communication dans laquelle il est inséré [9]. Cette notion de situation de communication est très importante pour la détermination de sa réalisation. Dans le contexte de la phonétisation d'un dictionnaire, il s'agit d'une situation de lecture de repérage phonétique de mots isolés. Dans ce cadre, un plus grand nombre de [ə] sont maintenus, particulièrement en syllabe initiale de mot et en syllabes contiguës. Dans un dictionnaire, au contraire, plusieurs transcriptions sont proposées pour un même mot et ces variations correspondent à différents contextes de la chaîne parlée, inexistants dans un lexique de mots orthographiques. Dans le cadre de l'effacement, réalisation [–], nous noterons "*" le "e", aussi bien dans la chaîne orthographique que dans la chaîne phonétique, pour garder la trace de son effacement

6. HARMONISATION DES RÉALISATIONS [±]

6.1. [ə] précédé de n consonnes

Dans le PR 90, la réalisation variable du [ə] n'est acceptée que dans les mots composés, avec ou sans tiret, à condition que le groupe consonantique précédant le [ə] commence par la liquide [ʀ] et qu'il soit suivi de deux syllabes, *port(e)bagage, port(e)balai, port(e)bouteille.* Nous le transcrirons aussi [±] dans les mots à haute fréquence du type *port(e)feuille, vers(e)ment, charg(e)ment, gard(e)rie.*
Dans les autres cas, il sera systématiquement maintenu en syllabe graphique initiale et interne de mot. pour la raison déjà évoquée de situation de lecture de mot isolé.

En syllabe finale il s'efface puisqu'il n'est suivi d'aucun autre son en contexte de mot isolé, *tabl*, êtr*.*

6.2. [ə] précédé d'une consonne

En syllabe initiale de mot, le [ə] est obligatoirement maintenu dans les mots littéraires ou techniques, *celer, repu, bedeau.* Il est transcrit [±] dans des mots plus courants, *v(e)nir, l(e)ver* dans le PR 90.
Dans nos règles de phonétisation, il sera systématiquement réalisé [+] pour éviter un groupe de deux consonnes à l'initiale de mot, surtout si ce dernier est court, *venir, besoin.* De même, le maintien du [ə] en monosyllabe est obligatoire pour la reconnaissance du mot outil lorsqu'il isolé. On imagine mal, en effet, *"le"* prononcé [l]. Nina Catach opte pour le même choix [4].

En syllabe interne de mot, le [ə] peut être situé, dans un premier cas, en syllabe initiale de lexème, lui-même précédé d'un préfixe, *apesanteur, maintenu, préretraite.* Sa transcription est variable dans le PR 90. Nous transcrirons son alternance [±] seulement dans les mots très fréquents, *souv(e)nir, bienv(e)nue.* Dans les mots moins fréquents, il sera maintenu, *anavenin, dépeçage.*
Dans les autres cas, le [ə] en syllabe interne de mot, n'est pas réalisé, *ach*ter, bib*ron, bais*main.*

En fin de mot, il ne sera pas maintenu, suivi ou non d'un "s" graphique final, *jup*, parol*, clopinett*s, condoléanc*s.*

6.3. [ə] en syllabes contiguës

Il existe des suites de deux et de trois [ə]. Ces suites apparaissent soit en syllabe initiale soit à l'intérieur de mot. Le critère essentiel que nous avons retenu est, rappelons-le, celui de la notion de situation de communication de lecture de mot isolé.

6.3.1.[ə] en syllabes 1 et 2

Dans le PR 90 le [ə] en syllabe initiale a toujours été maintenu sauf dans certains mots contenant le préfixe *"re"* où deux transcriptions ont été proposées, *red*mander, r*demander.*

Pour éviter un groupe de deux consonnes à l'initiale de mot tous les [ə] en syllabe 1 seront maintenus.

Le [ə] en syllabe 2 a été systématiquement trancrit [±], dans le PR 90, lorsque celui en syllabe initiale était maintenu. Nous opterons aussi pour son alternance, *rel(e)ver, ret(e)nir.*

6.3.2. [ə] en syllabes 2 et 3

Les transcriptions adoptées par le PR 90 sont variables, *échev(e)lé, irr*cevabilité, ensev*lir, brev(e)ter.* Comme la plupart des mots contenant cette suite contiennent un préfixe, le [ə] se trouve donc en syllabe initiale de lexème. C'est pourquoi, pour une meilleure lisibilité, nous choisissons de le maintenir en syllabe 2 et de transcrire [±] celui en syllabe 3. Ce choix correspond aux réalisations du [ə] que nous avons obtenues dans nos contrôles, *bêchev(et)er, échev(e)lé, irrec(e)vabilité, ensev(e)lir.*

Il existe quelques rares exceptions, *can*petière, gob*letier*, dans lesquelles le [ə] en syllabe 3 se réalise [+] vraisemblablement à cause de la présence du [j] dans le groupe consonantique qui le suit. L'alternance effacement / maintien est alors inversée. Enfin, dans le mot *bill*vesée*, c'est pour des raisons étymologiques que le [ə] en syllabe 2 s'efface : on peut le considérer comme un mot composé et le "e" en syllabe 1 se réalise [–] comme en finale de mot.

6.3.3. [ə] en syllabes 1. 2 et 3

Il existe seulement deux mots contenant une suite de trois [ə] transcrits par le PR 90, de la manière suivante : *red*venir ou r*dev*nir* et *r(e)ssem*ler.* Pour les mêmes raisons que pour les suites de deux [ə], le premier sera maintenu. D'après nos contrôles, le deuxième est réalisé [+] et le troisième s'est effacé : *redev*nir, ressem*ler.* C'est cette transcription que nous adopterons.

7. CONCLUSION

La phonétisation du "e" pose le problème plus général des variations phonétiques non seulement vocaliques mais aussi consonantiques. Dans le cadre de la transcription d'un français marqué, sans accent régional repérable, et représentatif, nous avons tenté d'être fidèle à la richesse de variations de réalisations tout en respectant une norme du français contemporain correspondant au français parlé de la fin du XXe siècle PR 90. Nous avons, en outre, lors de ce travail, collaboré avec Aliette Boumendil, chargée de la transcription phonétique du Petit Robert, qui, elle aussi, nous a servi de référence. A la lumière des discussions fructueuses que nous avons eues et des contrôles auxquels nous avons procédé, certaines notations phonétiques seront modifiées dans l'édition 1993 de ce dictionnaire.

Nous avons également essayé, dans l'écriture des règles de phonétisation, de traduire une image exacte du système grapho-phonétique du français en évitant ainsi que l'aspect technique ne l'emporte sur l'aspect linguistique.

8. RÉFÉRENCES

[1] AUBERGÉ, V., BELRHALI, R., BOË, L.J. & LIBERT, L. (1992). Organisation des lexiques pour une grammaire de phonétisation du français. Actes du séminaire Lexique. G.D.R. P.R.C. Communication Homme-Machine. IRIT - UPS. Toulouse.

[2] BELRHALI, R. & LIBERT, L. (1991). Phonétisation automatique : évaluation et enrichissement d'une grammaire de phonétisation du français. Mémoire de D.E.A. Université Stendhal. Grenoble.

[3] BLANCHE BENVENISTE, Cl., & CHERVEL, A. (1969). L'orthographe. Maspero. Paris.

[4] CATACH, N. (1984). La phonétisation automatique du français. C.N.R.S. Paris.

[5] CATACH, N. (1986). L'orthographe française. Nathan. Paris.

[6] DELL, F. (1985). Les règles et les sons. Hermann. Paris.

[7] LACHERET-DUJOUR, A. (1991). Le débit de parole : un filtre utilisé pour la génération des variantes de prononciation en français Parisien. Actes du XIIe Congrès International des Sciences Phonétiques. Aix en Provence, 3, 194-197.

[8] POTTIER, B. (1985). Linguistique générale. Klincksieck. Paris.

[9] RICHETERICH, R. (1976). Les situations de communication et les types de discours. Le Français dans le Monde, 121, 30-35.

[10] REY-DEBOVE, J. (1990). Le Petit Robert. Editions Le Petit Robert. Paris.

[11] VAN EIBERGEN, J. (1986). Le E latent. Bulletin de l'Institut de Phonétique de Grenoble, 15, 75-107.

[12] VAN EIBERGEN, J. (1991). Evaluation quantitative de l'alternance phonétique du /ə/. Importance de l'entourage consonantique. Actes du XIIe Congrès International des Sciences Phonétiques. Aix en Provence, 2, 150-152.

[13] VAN EIBERGEN, J. (à paraître, 1992). Application des règles phonologiques du /ə/ proposées par F. Dell. Les Cahiers de l'I.C.P. Bulletin de la Communication Parlée.

[14] WALTER, H. (1976). La dynamique des phonèmes dans le lexique français contemporain. France Expansion. Paris.

19es J.E.P. - Bruxelles, 19 au 22 mai 1992.

UN INVENTAIRE DE MOUVEMENTS MÉLODIQUES EN FRANCAIS

Frédéric Beaugendre † Christophe d'Alessandro † Anne Lacheret-Dujour †‡
Jacques Terken §
†LIMSI-CNRS, BP 133, F91403 Orsay Cedex
‡Univ. de Caen, Départ. de Linguistique Française, Place de la Paix F14000 Caen
§Institut de Recherche en Perception/ IPO, P.O. Box 513, NL5600 MB Eindhoven

RESUME

Cet article décrit un modèle intonatif pour la synthèse à partir du texte en français. Nous présentons ici les premières étapes de ce développement: l'étude acoustique et perceptive des courbes intonatives dans un corpus de 220 phrases isolées, lues par 3 locuteurs parisiens. Pour développer ce modèle, la méthodologie proposée par l'Institut de Recherche en Perception (IPO) a été suivie. Dans un premier temps les courbes mélodiques du corpus ont été stylisées pour obtenir des copies exactes en termes de mouvement mélodiques élémentaires. Dans un second temps, les mouvements mélodiques ont été classés en 9 mouvements standard. L'évaluation perceptive est discutée à chaque étape.

1 Introduction

Cet article présente une étude acoustique et perceptive d'un modèle mélodique pour la synthèse à partir du texte en Français. L'approche adoptée suit la méthodologie développée à l'Institut de Recherche en Perception (IPO, Eindhoven, Pays-Bas) depuis une trentaine d'année, méthodologie qui à été appliquée avec succès à l'étude intonative du Hollandais [7], du Russe [12], de l'Allemand [1] et de l'Anglais Britannique [14].

L'application de cette méthodologie n'a pas jusqu'à présent été vraiment réalisée pour le Français, lacune que ce travail vise à combler. Il faut toutefois mentioner les travaux menés au MIT par Vaissière [15,16] et Delgutte [3], bien que ces travaux n'aient pas exploré l'aspect perceptif de l'intonation (voir [17] p.545 "Dans notre cas il ne s'agit pas de mouvements intonatifs (en rapport avec la perception) mais de mouvements du fondamental, car notre étude est basée avant tout sur une étude acoustique du signal").

L'originalité de la démarche d'IPO est de prendre le signal acoustique comme point de départ de l'étude intonative, et de n'accepter des hypothèses prosodiques fonctionnelles que dans la mesure où elles apparaissent vérifiables, c'est à dire audibles à travers des manipulations systématiques de courbes intonatives synthétiques.

L'analyse qui suit s'attache à la description de *l'intonation*, c'est à dire la variation de la fréquence fondamentale (F0) perçue au cours du temps. D'autres paramètres doivent être pris en compte, en particulier la structure syllabique et le contenu segmental de l'énoncé. Il s'agit de caractériser les contraintes acoustiques et perceptives liées aux courbes mélodiques, sans considérer au départ leur fonction linguistique. La resynthèse et l'évaluation perceptive sont les instruments de cette caractérisation acoustiço-phonétique. Contrairement aux critiques fonctionnalistes (voir par exemple [11]) nous pensons que cette hypothèse de travail est complémentaire des approches linguistiques de l'intonation.

Les étapes pour la génération automatique de l'intonation dans un système de synthèse à partir du texte sont:

1. la récolte d'un corpus de signal construit en accord avec les contraintes liées à notre propos;
2. l'analyse, la resynthèse et la stylisation des courbes intonatives, à partir d'une évaluation perceptive;
3. la construction d'un *modèle mélodique*, c'est à dire un ensemble de mouvements intonatifs élémentaires qui permettent perceptivement de reconstruire les courbes stylisées;
4. le développement d'une *grammaire de l'intonation*, spécifiant les règles d'enchaînements licites des mouvements en contours mélodiques;
5. la mise au point d'un inventaire de patrons intonatifs pour rendre compte de l'organisation des contours mélodiques syntaxiquement corrects en termes fonctionnels;
6. la construction de règles pour passer de la structure (linguistique, rythmique) de l'énoncé à sa structure intonative;

7. l'analyse du texte à synthétiser pour obtenir sa structure rythmique, syntaxique et accentuelle.

Seuls les trois premiers points seront abordés ici. La récolte du corpus de texte et de signal sont présentés dans la première partie. La seconde partie est consacrée à la méthodologie d'analyse et de stylisation des courbes intonatives, ainsi qu'au logiciel de stylisation développé, au corpus stylisé et son évaluation perceptive. La troisième partie traite de la constitution du modèle mélodique: la classification des mouvements intonatifs en catégories de mouvements prototypes. Les critères acoustiques et/ou perceptif choisis, et l'évaluation perceptive des courbes synthétiques obtenues sont présentés.

2 Corpus et analyse

2.1 Corpus de texte

La tâche fixée était de construire un corpus de texte et de signal pour la synthèse à partir du texte.

Un corpus de 220 phrases a été constitué en fonction de contraintes syntaxiques, lexicales, phonétiques et phonotactiques. Les contraintes syntaxiques concernent la modalité de la phrase (assertive, interrogative, négative, impérative) et sa structure (phrase simple ou phrase complexe).Les contraintes lexicales sont liées à la nature du mot (mot simple, mot dérivé ou mot composé) et à sa fréquence d'utilisation relative (mot rare, mot courant). Les contraintes phonétiques interdisent dans la mesure du possible la présence d'occlusives sourdes à l'initiale de mots. Enfin, les contraintes phonotactiques sont relatives à la position du mot dans la phrase et à sa longueur (nombre de syllabes).

2.2 Récolte du signal

Ce corpus de phrases isolées a été lu en cabine insonorisée par trois locuteurs parisiens a priori représentatifs d'un français non marqué (1F, 2H). Un des locuteurs masculins a également participé à l'enregistrement d'un corpus pour l'élaboration des règles acoustico-phonétiques utilisées dans le système de synthèse à partir du texte en cours de développement au LIMSI [6]. De par la situation de lecture imposée, la prononciation est appliquée et le style prosodique résultant peut-être qualifié de "neutre": les accents emphatiques ou émotionels sont généralement absents. Le signal a été recueilli par un microphone B&K4165, et numérisé directement à une fréquence d'échantillonage de 16kHz.

2.3 Analyse du signal

Le signal récolté a été analysé par une LPC30, une détection automatique de F0 avec l'algorithme AMDF a été effectuée. Les erreurs de l'analyse de F0 ont été corrigées manuellement afin d'obtenir une resynthèse correcte du point de vue intonatif. Nous avons choisi une analyse LPC malgré la dégradation de qualité du signal de resynthèse par rapport au signal original. Car, d'une part cette dégradation n'altère pas significativement la perception de l'intonation, d'autre part la simplification extrême de l'excitation voisée du codage LPC classique autorise une modification très précise de l'intonation. Des tests comparatifs entre un système de codage LPC et un système de modification de F0 par découpage/collage de périodes de voisement ont été effectués. Ces tests ont montré que, lorsque l'on analyse à nouveau le fondamental d'un signal resynthétisé par chacun des systèmes, on retrouve beaucoup plus exactement la mélodie choisie dans le cas de la LPC. Le signal obtenu a été étiqueté phonétiquement, avec un marquage des coupes syllabiques.

3 Stylisation des courbes intonatives

3.1 Méthode de stylisation

La procédure de stylisation vise à réduire les courbes intonatives naturelles en une concaténation de segments de droite, dénomés *mouvements mélodiques.* Les mouvements mélodiques sont des segments de droite dans le plan temps/F0 suivant des axes exprimés en seconde/demi-ton. Le choix d'une unité de type logarithmique pour représenter les fréquences permet de tenir compte de notre perception mélodique, qui n'est pas linéaire. Les discontinuités de la dérivée des courbes F0 aux points de raccordement ne semblent pas perceptivement pertinentes. A ce niveau d'analyse, le but est d'obtenir des *copies exactes* (close copy stylisation [14]) à partir des *originaux synthétiques* (c-à-d du codage LPC du signal original). Plusieurs copies exactes peuvent être établies pour le même original synthétique, mais si l'on impose comme contraintes que:

1. les copies exactes doivent être perceptivement indiscernables de l'original synthétique;
2. le nombre de mouvements mélodiques d'une copie exacte doit être minimum;

l'expérience montre que finalement assez peu de possibilités significativement différentes subsistent. En outre, l'étape suivante de la méthode doit pouvoir associer dans une même catégorie des mouvements mélodiques voisins. Il ne faut cependant

pas sous-estimer la difficulté d'automatiser la stylisation. La non-unicité des stylisations provient de l'expertise du stylisateur, qui utilise de nombreuses boucles d'analyse par la synthèse. Cette expertise semble difficilement formalisable. La stylisation est effectuée en utilisant le logiciel décrit dans le paragraphe suivant.

3.2 Environnement de stylisation

Le logiciel de stylisation à été developpé pour compatible IBM PC, équipé d'une carte de traitement de signal OROS ou VECSYS. L'écran de travail contient trois fenêtres. La première présente l'affichage de la courbe mélodique et la stylisation en respectant une echelle demi-ton/seconde. La deuxième est dédiée à l'affichage du signal (amplitude/seconde). La troisième fait apparaître la représentation phonétique du texte, préalablement segmenté à l'aide d'un spectrogramme. L'affichage de cette fenêtre n'est pas validé pendant l'étape de stylisation afin de ne pas faire interférer niveau segmental et niveau suprasegmental.

La stylisation est réalisée à l'aide de la souris, avec laquelle on indique les extrémités de chacun des segments.

L'opérateur peut à tout moment écouter l'énoncé original, l'original resynthétisé ou l'énoncé stylisé sur la phrase entière ou une portion selectionnée.

Une étape préalable à la stylisation consiste à corriger les éventuelles erreurs de détection du fondamental, notamment aux frontières de voisement, à l'aide d'une fonction également disponible dans cet environnement graphique.

Il est enfin possible d'imprimer une copie d'écran (voir figure 1).

3.3 Evaluation perceptive des copies exactes

La stylisation suppose les hypothèses suivantes:

1. on peut procéder à une réduction des données sur la base de tolérances perceptives.

2. on distingue deux types de variations de la fréquence fondamentale. Le premier (macroprosodique) est volontairement programmé par le locuteur. Le second (microprosodique) rend compte au contraire des variations involontaires de f0 dues à l'articulation. Ces variations n'apportent pas une contribution essentielle à la perception de l'intonation, elles peuvent donc être omises.

La première proposition est vérifiable expérimentalement. Des tests informels pour le français montrent que copies exactes et originaux synthétiques, sont effectivement indiscernables au niveau de la phrase, si les copies ont été soigneusement réalisées. Cependant, lorsque le contexte de comparaison est réduit à une ou quelques syllabes, des différences peuvent être systématiquement perçues par des sujets entraînés à l'écoute fine des hauteurs. Des tests utilisant un paradigme identique/différent pour trois types de stimuli (copies exactes, originaux synthétiques, copies avec déplacement d'un accent) sont reportés dans [14] pour l'anglais britannique. Les résultats montrent que les copies exactes sont systématiquement confondues avec les originaux synthétiques, et les copies modifiées systématiquement distinguées. Pour affiner l'évaluation de l'identité perceptive entre copies exactes et originaux synthétiques suivant les restrictions apparues dans notre évaluation informelle, nous avons défini un protocole ABX (A: copie exacte, B: original synthétique, X: l'un ou l'autre) en variant le contexte d'une syllabe à une phrase. Les tests sont actuellement en cours.

La seconde proposition est plus problématique. En effet, la distinction de la microprosodie du reste des phénomènes intonatifs est délicate [12,11]. On ne peut donc pas prétendre ici à un traitement systématique de la microprosodie: parfois on peut éliminer les micro-fluctuations (donc tracer un segment à travers une consonne occlusive ou constrictive), mais parfois cette micro-fluctuation donne naissance à un changement de mouvements mélodiques. Néanmoins, il faut rappeler que les tests à cette étape relèvent d'un mode psycho-acoustique de perception (deux mélodies sont elles indiscernables pour des sujets entraînés dans des conditions de laboratoire ?), et non d'un mode phonétique (deux mélodies ont-elles la même prosodie ?). Pour le propos de représenter l'intonation comme phénomène prosodique, les tests sont clairement trop sévères à cette étape.

3.4 Discussion: mouvements mélodiques, commandes d'accent, valeurs cibles

Le propos de cette partie étant la copie de courbes intonatives naturelle par des courbes suivant un certain modèle, il faut discuter du choix du modèle. Dans le modèle de source vocale [5], le lien entre l'aspect fonctionnel et l'aspect acoustique est direct, puisque les commandes du modèle sont uniquement fonctionnelles (commandes d'accent, commande des groupes). Cependant, ce modèle, développé initialement pour le japonais, ne permet pas directement de rendre compte de toutes les variations intonatives relatives à la modalité d'un énoncé, la modalité interrogative notamment.

Une autre ligne de recherche modélise les variations des hauteurs mélodiques par l'interpolation

(polynomiale, ou par fonctions splines) de valeurs cibles [9] [13]. Il semble nécessaire d'ajouter à la description en valeurs cibles, liées à des *niveaux mélodiques* une dimension temporelle (l'instant relatif auquel le niveau est atteint). Les deux approches, valeurs cibles et mouvements mélodiques ne nous semblent alors pas fondamentalement incompatibles. Il est probable que la prise en compte d'un modèle plus fin de la perception mélodique, comme celui proposé dans [10] permettra d'unifier la description intonative au niveau purement acoustique.

4 Modèle mélodique

Un corpus d'environ 70 phrases a actuellement été stylisé suivant la procédure décrite dans la partie précédente. Ces copies exactes comportent environ 500 mouvements élémentaires. Dans cette partie nous décrivons la procédure de standardisation des mouvements mélodiques. Cette procédure vise à remplacer les mouvements obtenus dans les copies exactes par des mouvements prototypes. Les courbes mélodiques obtenues par concaténation de mouvements mélodiques standards sont dénomées *contours mélodiques*. Du point de vue perceptif, la relation recherchée entre copies exactes et originaux synthétiques est l'identité. Par contre, entre les contours mélodiques et les copies exactes, l'identité doit être remplacée par l'équivalence. Cette notion d'équivalence est plus difficile à définir: l'identité peut se mesurer à l'aide de critères psycho-acoustiques, mais l'équivalence doit être déterminée par une méthodologie psychophonétique. La compétence linguistique des sujets jugeant de l'équivalence des mélodies est explicitement solicitée, et l'équivalence perceptive nécessite un niveau d'abstraction plus élevé que l'identité perceptive. Deux courbes mélodiques seront dites équivalentes si:

1. elles sont perçues comme assez similaires du point de vue mélodique;
2. elle portent les mêmes fonctions linguistiques dans un contexte donné.

Pour le moment, l'évaluation perceptive de l'équivalence mélodique a été menée de façon informelle par les auteurs de cet article. Des tests formels, proposés dans [14,8], seront mis en oeuvre ultérieurement.

4.1 Critères de classification

La classification des mouvements mélodiques est soumise d'une part à la contrainte de l'équivalence perceptive et d'autre part à un principe d'économie: le minimum de mouvements standards doit être recherché. Ces mouvements diffèrent au moins par un trait acoustique. Les traits retenus pour la classification des mouvements sont :

1. la direction du mouvement (montée (R), descente (F) ou mouvement plat);
2. la pente du mouvement (pente forte, pente faible);
3. l'ambitus du mouvement (mouvement complet, demi-mouvement);
4. La durée du mouvement, liée aux durées syllabiques;
5. La position du mouvement par rapport à la syllabe. Précisons ici que les structures syllabiques rencontrées dans notre corpus sont du type "attaque consonantique et noyau vocalique" (CV) ou "noyau" seul (V).

4.2 Classification

La segmentation du continuum prosodique nous a amenés à distinguer initialement 13 mouvements. Cette première classification à été réalisée en ne considérant que la valeur des différents paramètres acoustiques, indépendamment de considérations perceptives.

Une analyse perceptive a ensuite permis de réduire le nombre de mouvements à 5 montées, 3 descentes et un mouvement plat. En effet, certaines variantes acoustiques, non pertinentes du point de vue perceptif, peuvent ne pas être prises en compte. S'il n'est pas question ici de rentrer dans les détails de tous les paramètres utilisés pour distinguer ces segments, nous pouvons néanmoins présenter les caractéristiques acoustiques fondamentales de ces différents mouvements. Les mouvements montants sont R1, que l'on trouve toujours sur une voyelle longue (durée supérieure à 100 ms), R2, mouvement court de pente forte localisé uniquement sur les consonnes, R3, mouvement couvrant toute une syllabe, enfin, R4 qui est un mouvement de durée variable et de pente faible, toujours enchaîné à un demi-mouvement montant de pente plus forte (R5). Les descentes sont F1, mouvement de pente faible, s'étalant sur plusieurs syllabes, F2, s'étalant sur une syllabe et de pente forte, enfin, F3, qui est un mouvement de pente forte situé sur une consonne. Le mouvement plat, de durée extrêmement variable, a pour seule fonction de relier les autres mouvements entre eux.

La dynamique de chacun de ces mouvements diminue au fur et à mesure que l'on avance dans la phrase. L'ambitus des mouvements aura été calculé

Table 1: Paramètres acoustiques des différentes classes de mouvements

	R1	R2	R3	R4	R5	F1	F2	F3
Direction								
montée	+	+	+	+	+			
descente						+	+	+
Timing								
attaque		+			+			+
noyau	+							
att-noyau			+				+	
Pente								
forte	+	+	+		+		+	+
faible				+		+		
Ambitus								
plein	+	+	+	+		+	+	+
demi					+			

en demi-tons, définis comme suit :

$$Ambitus = \frac{12}{\log 2} \times \log(\frac{f2}{f1})$$

avec : Ambitus = Différence de fréquence en demi-ton (DT)
f1 = fréquence initiale du mouvement mesuré
f2 = fréquence finale du mouvement mesuré

La valeur moyenne de chacun des mouvements est de 6 DT mis à part pour R4 et R5 (3 DT).

A partir de cette classification acoustique et perceptive, les premières constatations ont été faites quant à l'aspect fonctionnel des variations intonatives : le mouvement R1 correspond toujours à une fin de mot prosodique; R2 semble être perçu comme un changement de niveau intonatif et non comme un mouvement à proprement parler [10], au même titre que F2. Par ailleurs, il sera intéressant de comparer le modèle mélodique obtenu avec ceux d'autre langues, car on peut noter que certains mouvements ne sont pas spécifiques au français, ainsi F1, mouvement descendant graduel, est également rencontré pour le hollandais et l'anglais par exemple.

5 Conclusion

Nous avons présenté les premières étapes du développement d'un modèle intonatif pour la synthèse à partir du texte en français. A court terme, nos perspectives sont le développement d'une grammaire d'enchaînement de ces mouvements mélodiques, et la définition de patrons intonatifs en relation avec des critères syntaxiques. A moyen terme, l'étude systématique des variations mélodiques en relation avec la structure rythmique doit être envisagée. Enfin, à long terme une étude sur la variabilité inter-locuteur des stratégies intonatives est nécessaire dans le cadre des recherches sur la synthèse multi-voix et multi-style. Ce travail est mené dans le cadre du projet ESPRIT-POLYGLOT 2104.

References

[1] Adriens L. **(1991) Strukturen deutcher Intonation.** Doctoral dissertation, Eindhoven University of Technology.

[2] Collier R. **(1991) Multi-langage intonation synthesis.** Journal of Phonetics **19**, pp 61-73.

[3] Delgutte B. **(1976) Fundamental frequency contours of French: a perceptual study**, Msc Thesis, M.I.T.

[4] Emerard F. **(1977) Synthèse par diphones et traitement de la prosodie.** Thèse de $3^{ème}$ cycle, Université des Langues & Lettres de Grenoble.

[5] Fujisaki H. & Sudo H. **A Generative Model for the Prosody of Connected Speech in Japanese.**, IEEE Int. conf. an ASSP, Newton, pp. 140-143.

[6] Garnier-Rizet M. **(1991) A Rule-based Segmental Synthesis module for French.**, Eurospeech 91, Genova, **1**, pp. 51-54.

[7] 't Hart J., Collier R. & Cohen A. **(1991) A perceptual study of intonation : an experimental-phonetic approach to speech melody.** Cambridge University Press.

[8] 't Hart J. **(1991) F0 stylisation in speech : Straight lines versus parabolas.** J. Acoust. Soc. Am. **90** (6), pp 3368-3370.

[9] Hirst D.J. **(1977) Intonative Features.**, Mouton, Janua Lingarum, The Hague.

[10] House D. **(1991) A Model of optimal tonal feature perception.** 12th Int. Cong. Phon Sc. ICPhS, Aix en Provence, **2**, pp 102-105.

[11] Kohler K.J. **(1991) Prosody in speech synthesis: the interplay between basic research and TTS application.** Journal of phonetics **19**, pp 121-138.

[12] Odé C. **(1989) Russian Intonation : A Perceptual Description.** Amsterdam/ Atlanta GA : Rodopi.

[13] Pierrehumbert J. **(1981) Synthesizing Intonation.**, J. Acoust. Soc. Am., **70**, pp. 985-995.

[14] de Pijper J.R. **(1983) Modelling British English intonation.** Foris Publications, Dordrecht.

[15] Vaissière J. **(1974) On french prosody.**, Res. Lab. Electr., Q. Prog. Report No.115, M.I.T., 251-262.

[16] Vaissière J. **(1975) Further Note on French Prosody**, Res. Lab. Electr., Q. Prog. Report No 115, M.I.T., 251-262.

[17] Vaissière J. **(1980) La structuration acoustique de la phrase française**, Annali della scuola normale superiore di Pisa, pp. 530-560.

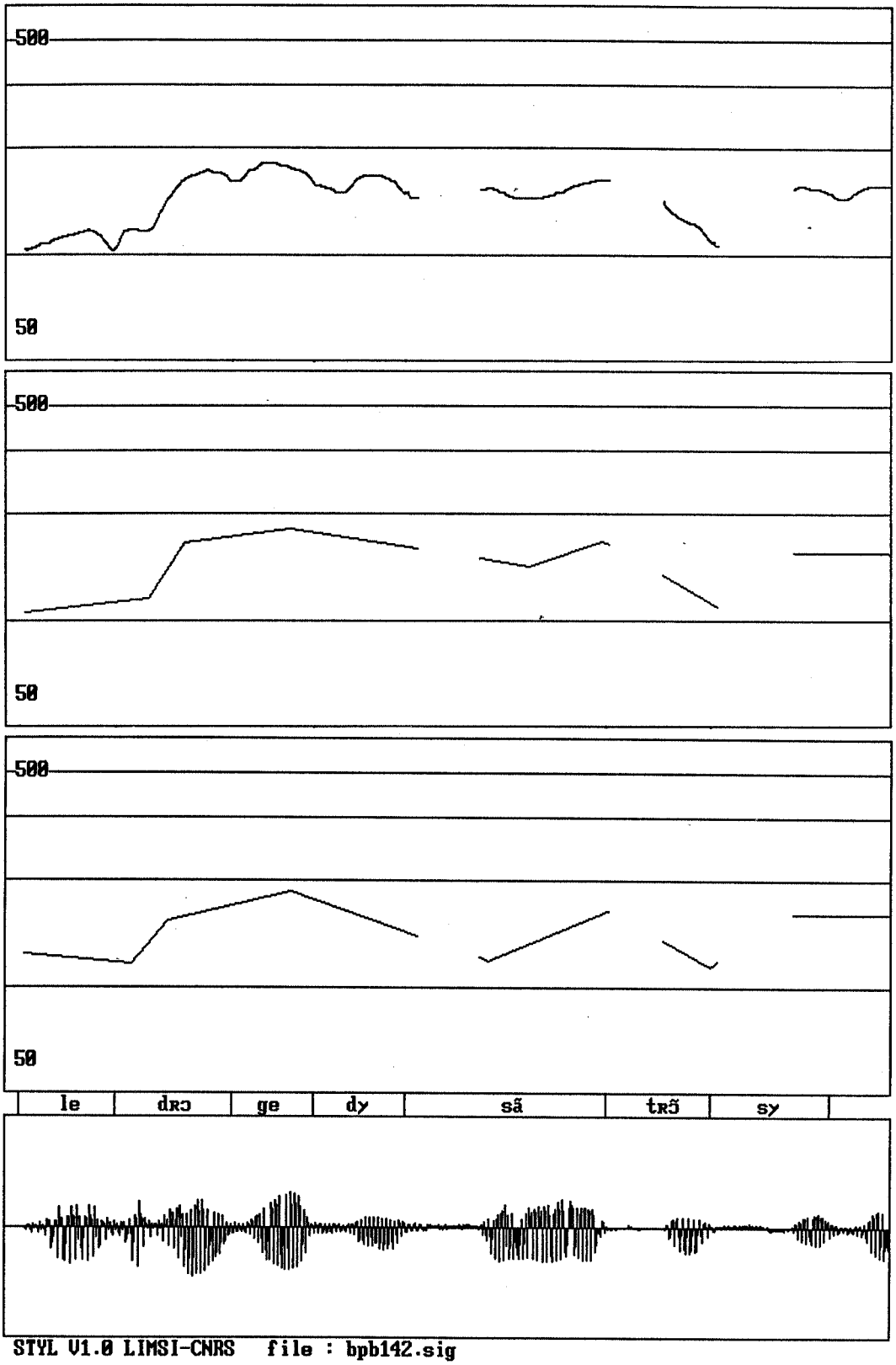


Figure 1: Courbe intonative naturelle (A), copie exacte (B), contour mélodique avec des mouvements standards (C)

19es J.E.P. - Bruxelles, 19 au 22 mai 1992.

GENERATION AUTOMATIQUE DES «P-CENTERS»

BARBOSA, P. et BAILLY, G.

**INSTITUT DE LA COMMUNICATION PARLEE,URA CNRS N° 368,
INPG - UNIVERSITE STENDHAL
46, AV. FELIX VIALLET- 38031 - GRENOBLE CEDEX - FRANCE**

Résumé

Nous présentons ici un modèle de prédiction de la durée segmentale par apprentissage automatique qui procède en deux étapes: une prédiction de l'intervalle inter «P-Centers» par réseau séquentiel suivi d'une répartition de cette durée entre les divers constituants de l'unité au moyen d'une méthode statistique.

1. INTRODUCTION

Les premiers modèles de prédiction automatique de la prosodie pour la synthèse à partir du texte étaient basés sur une description phonologique du continuum prosodique. L'*analyse statistique* comme la *modélisation stochastique* permet d'adapter un tel squelette à la réalisation effective d'un locuteur donné dans une situation donnée [Emerard, 77; Traber, à paraître; Aubergé, à paraître; Sagisaka, 90]. Ces systèmes permettent d'associer à chaque phonème un nombre constant de valeurs de paramètres prosodiques en fonction de leur contexte et de la nature de l'unité phonologique à laquelle ils appartiennent.

La conception de modèles stochastiques pour la prédiction de la fréquence fondamentale est facilitée par la relation de la courbe mélodique aux constituants phonématiques: chaque syllabe est caractérisée par un nombre constant de cibles (typiquement de 1 à 3) dont les valeurs sont considérées comme indépendantes de la nature des constituants (la microprosodie étant modélisée séparément). Cette dernière hypothèse ne peut pas être évidemment retenue pour la durée segmentale: celle-ci dépend étroitement de la nature intrinsèque du segment ainsi que de son contexte [Di Cristo, 85; Lehiste, 77]. Ceci explique pourquoi l'approche statistique (bien décrite par van Santen & Olive, 90) est relativement systématique dans ce domaine de recherches.

L'approche modèle [Todd, 89] a pourtant les avantages classiques d'interpolation et de lissage qui permettent de pallier aux insuffisances des corpus d'apprentissage et de résoudre les problèmes d'interpolation de contours à la synthèse. Elle permet aussi de contraindre l'apprentissage par la structure même du prédicteur et non pas seulement par la structure des données.

Notre travail consiste à élaborer une structure de prédicteur multiparamétrique de la prosodie basé sur la notion de contrôle gestuel [Bailly *et al.*, 91]: les unités phonologiques décrivent la tâche et le contrôleur l'exécute en fonction de contraintes rythmiques propres à la langue.

2. DE LA NATURE DE L'UNITE DE PROGRAMMATION RYTHMIQUE

Le mouvement quasi-périodique de la mâchoire a très tôt mis en évidence le rôle de la syllabe dans l'organisation rythmique de la parole [Fraisse, 74] tant au point de vue de sa perception que de sa production.

Le travail mené par Marcus autour des «Perceptual Centers» (PC) montre que des auditeurs sont capables de régler de manière

consistante le retard entre deux récurrences syllabiques (V versus CV) de manière à percevoir cette alternance comme isochrone [Marcus, 76]. Cette tendance à l'isochronisme semble être une des composantes du système de contrôle rythmique ainsi que l'atteste le débat sur l'isochronisme et l'isosyllabisme [Pike, 45; Lehiste, 77]. Le PC est considéré comme l'instant dans la syllabe de test CV à égale distance des établissements de la voyelle de référence V [Pompino-Marschall, 89]. Malgré la diversité des consonnes C, le PC semble être dans le voisinage immédiat de l'établissement du noyau vocalique.

Dans la suite, nous allons envisager les candidatures de la syllabe et de l'intervalle inter-PC comme unités de programmation rythmique.

2.1. LE SYSTEME DE PRODUCTION PROPOSE PAR CAMPBELL

Prenant comme hypothèse que la syllabe est l'unité de programmation rythmique, Campbell propose un modèle de génération de la durée segmentale pour l'anglais en deux étapes [Campbell, à paraître]:

• génération de la durée syllabique qui ne dépend que de la typologie accentuelle, du nombre de constituants de cette unité et de la nature du noyau syllabique;

• répartition de cette durée entre ces constituants selon un modèle statistique .

La première étape de cette prédiction se prête bien à un apprentissage automatique de part la faible dimension de son entrée (6 paramètres par syllabe). Cette prédiction est effectuée par un réseau multi-couches.

Le calcul de la durée segmentale Duri se fait au moyen d'un facteur de déformation k appliqué à la moyenne μi et à l'écart-type σi des durées phonémiques en millisecondes log-transformées suivant la formule: $Duri = \exp(\mu i + k.\sigma i)$, avec k tel que Σ Duri = Durée de la syllabe.

L'analyse de la valeur moyenne des facteurs k de quatre types de segments («onset», «peak», «coda», «medial») montre la bonne homogénéité des facteurs k de ces segments malgré les différences considérables des caractéristiques temporelles des consonnes et voyelles. Il est cependant à noter chez Campbell que le «coda» des syllabes longues et l'«onset» (établissement) des syllabes finales ont un coefficient assez différent des autres constituants.

La même analyse effectuée pour le français sur le corpus décrit ci-dessous est présentée Table 1.

TABLE.1. Moyennes, écarts-type et nombre d'occurences des facteurs k pour 3 types de syllabes.

	syllabes longues			syllabes courtes			syllabes finales		
	moy.	éc.type	n	moy.	éc.type	n	moy.	éc.type	n
établiss.	1,36	0,39	120	-1,55	0,40	283	-0,30	0,44	9
tenue	1,45	0,42	221	-1,60	0,40	145	1,79	0,46	49
coda	1,41	0,36	12	-1,39	0,64	17	2,85	0,84	21

Les tests d'équivalence des moyennes effectués nous a permis de garder l'hypothèse nulle pour les syllabes longues et courtes tandis qu'elle a été rejetée pour les syllabes finales: les différences entre moyennes ne peuvent être interprétées de manière significative et il faut donc avoir recours à une analyse plus fine permettant de faire émerger ce qui se passe localement.

2.2. APPLICATION DU MODELE DE REPARTITION AU FRANÇAIS

2.2.1. BASE DE DONNEES

Deux corpus ont été utilisés :

• Un corpus de plus de 2000 logatomes enregistrés à débit normal sur lequel on a fait les statistiques des durées log-transformées des réalisations phonémiques. Le résultat se trouve en annexe.

• Le corpus dont les durées ont été prédites contient 88 phrases enregistrées à débit normal et rapide. Ce corpus a été conçu de manière à mettre le locuteur en situation de délivrance d'information (type: «Rappelez M. Dupont à son bureau !») et les phrases possèdent entre 4 et 33 syllabes.

2.2.2. ETUDE DES CORRELATIONS ENTRE LES FACTEURS k ENTRE SEGMENTS

Nous avons cherché par une étude des corrélations quelle était la validité de l'hypothèse de travail de Campbell, soit le comportement uniforme des déformations des divers segments. Pour cela, les constituants de la syllabe classique ont été classés en *établissement* (la charge consonantique précédant la voyelle), *tenue* (la voyelle) et *coda* (la charge consonantique la suivant). Pour cette analyse les syllabes finales ont été mises à part ainsi que suggéré par Campbell

La corrélation concerne un segment courant et celui qui le suit: cela nous permettra de défendre soit une approche syllabique soit de proposer une autre unité pour le geste rythmique. Les résultats sont donnés table 3. Il n'y avait pas de paire *coda/coda* dans notre corpus (groupe consonantique final).

TABLE.2. Corrélation par méthode de Pearson entre segments adjacents. Les coefficients de confiance sont données entre parenthèses.

segment suivant ->		établiss.	tenue	coda
segment	établiss.	0,51 (p<<0,001)	**-0,20** (p<<0,001)	-
courant	tenue	**0,12** (p<<0,001)	-	0,17 (p<0,1)
->	coda	-0,13 (p<0,2)	-0,41 (p<0,005)	-

Pour les paires *tenue / établissement* appartenant au même groupe PC on a trouvé une corrélation de 0,12 et pour les paires *établissement / tenue* (même syllabe), la valeur a été - 0,20.

Cela nous conduit à envisager un groupe du type PC entre deux établissements vocaliques consécutifs. Ainsi, si l'on considère l'ensemble des paires pour une approche syllabe classique (établiss./établiss., établiss./tenue, tenue/coda), on trouve r=-0,00 alors que pour une approche PC (tenue/établiss., établiss./établiss., coda/établiss., tenue/coda), on trouve r=+0,26.

3. GENERATION DE LA DUREE SEGMENTALE

Les principales hypothèses de notre modèle sont:

• les diverses durées inter «P-centers» sont obtenues par la déformation d'une horloge interne [Allen, 75]. Cette horloge biologique serait le point de repère de l'isochronisme et le déclencheur des actions motrices [Turvey *et al*, 90]. Cette horloge interne constitue un attracteur rythmique: le mouvement accentuel consisterait à indiquer «qu'il s'est passé quelque chose» par le déphasage des deux horloges [Lehiste, 77] qui tomberaient en phase suite à la réalisation de l'accent [Barbosa, 91].

• les durées segmentales peuvent être obtenues statistiquement à partir des distributions de chaque classe de son, tout en considérant que la valeur de corrélation présentée ci-dessus n'est pas négligeable.

3.1 GENERATION DE LA DUREE INTER-PC PAR RESEAU SEQUENTIEL

Vu la difficulté de bien cerner les facteurs influençant la durée des groupes inter «P-centers», on a mis en place une architecture connexionniste du type séquentiel [Jordan, 86] représentée ci-dessous.

L'entrée du réseau contient la description phonologique, caractérisée par la marque prosodique [Bailly, 89] courante et suivante, la nature de la tenue courante et suivante, la charge consonantique (nombre de consonnes inter-PC). Des rampes négatives générées sur chaque groupe accentuel ainsi que sur la phrase indiquent la position du groupe inter-PC courant dans l'unité concernée. Cette description est actualisée à chaque période de l'horloge interne. Le nombre total de périodes par phrase est égal au nombre de groupes inter-PC.

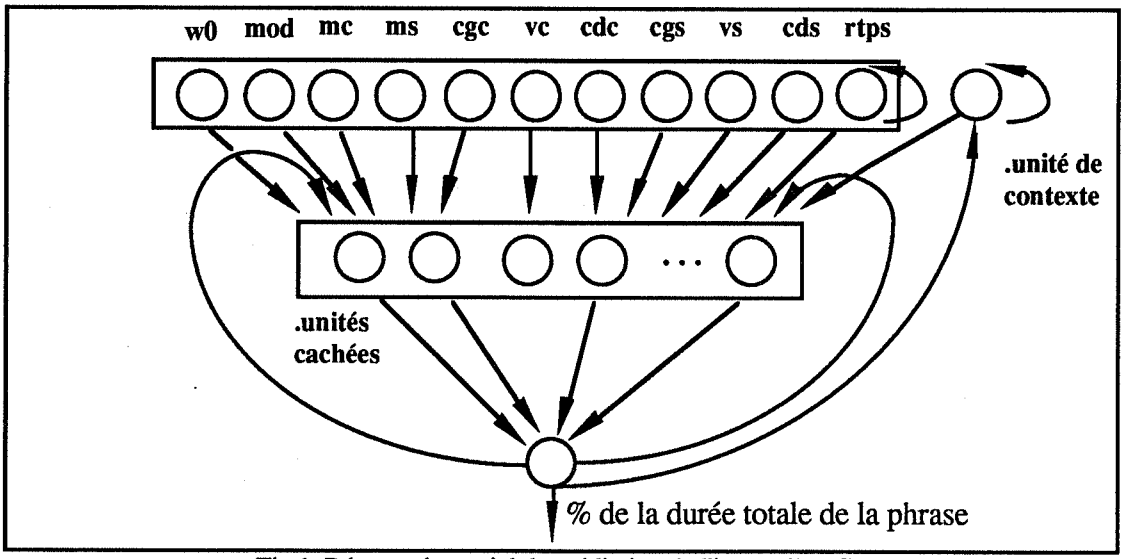


Fig 1. Réseau séquentiel de prédiction de l'intervalle PC.

L'unité de sortie est la durée du groupe inter «P-centers» courant. Celle-ci a été codée en pourcentage par rapport à la période totale entre le premier et le dernier PC. La distribution des syllabes ainsi codées s'approche bien de la normale.

La moitié du corpus a été utilisée pour l'apprentissage, l'autre moitié, pour le test. L'erreur totale de prédiction du réseau pour les phrases de test a été inférieure à 1,3 fois celle du corpus d'apprentissage.

3.2 LA DISTRIBUTION DE LA DUREE ENTRE LES CONSTITUANTS

La sortie du réseau présenté précédemment a servi d'entrée pour l'algorithme de distribution de la durée du groupe entre ses constituants.

A titre d'exemple (pour simplifier on suppose les moyennes et écarts-types exprimés en ms), on prend les groupes /as/ et /aʀ/. On a les moyennes $\mu a = 114$ ms, $\sigma a = 40$ ms; $\mu s = 166$ ms, $\sigma s = 48$ ms; $\mu R = 94$ ms, $\sigma R = 39$ ms. Pour une durée totale de 200 ms, on obtient pour /as/, Dura=78 ms, Durs=122 ms (k=-0.91) alors que pour /aʀ/, Dura=110 ms, DurR=90 ms (k=-0.1). Cette simulation montre que le /a/ sera plus long de 35 % s'il est suivi d'un /ʀ/ que s'il est suivi d'un /s/. Ce chiffre est à rapprocher des fameux coefficients de correction «co-intrinsèques» de la durée [Di Cristo, 85; Bartkova & Sorin, 87].

Cette méthode a présenté comme erreur 18 ms par segment pour les vraies durées inter-PC, soit 20% de la moyenne de durée segmentale du corpus ou 25 ms par segment pour les durées prédites par le réseau (corpus d'apprentissage et de test confondus), soit 28% de la moyenne.

4.CONCLUSIONS ET PERSPECTIVES

L'étude des facteurs k nous a montrés la corrélation systématique plus forte d'un groupe consonantique intra-syllabique homogène (établissement ou coda) avec la voyelle précédente, ce qui conforte l'approche PC présentée ici. Néanmoins, l'étude des transitions *coda / établissement* ($r = -0{,}13$, $p < 0{,}2$) suggère que le PC pourrait dans ce cas être déplacé du début de la tenue à cette frontière. Mais il reste à entreprendre des expériences de perception dans cette configuration. En tout cas, les corrélations faibles montrent qu'il y a des non-linéarités dans la répartition de la durée entre les segments.

Bien que les modèles connexionnistes permettent de mettre en évidence les paramètres de commande pertinents, la convergence lente du réseau peut être le signe d'une redondance dans la description phonologique, de la non pertinence de certains facteurs à déterminer la structuration rythmique de la phrase ou bien sûr du manque de finesse de la description phonologique.

Des tests pour le corpus à débit lent doivent être effectués pour intégrer le phénomène pausal dans cette génération rythmique (par exemple par un très grand coefficient d'élasticité de la pause associé au modèle non-linéaire de répartition des durées) et exploiter l'hypothèse que les horloges interne et déformée tombent en phase après la réalisation de l'accent énonciatif.

REMERCIEMENTS

Les corpus proviennent de l'ENST et nous remercions F. Bimbot de nous avoir prêté sa voix. Ils ont été étiquetés avec soin par notre ami R. Sock. Le simulateur de réseaux séquentiels a été conçu et réalisé par R. Laboissière.

REFERENCES BIBLIOGRAPHIQUES

Allen, G.D. (1975) "Speech rhythm: its relation to performance universals and articulatory timing", *Journal of Phonetics*, **3**, 75-86.

Aubergé, V. (à paraître) "Developing a structured lexicon for the synthesis of prosody", in *Talking machines : theories, models and designs* , (Bailly,G. & Benoît, C., Eds).

Bailly, G. (1989) "Integration of rhythmic and syntactic constraints in a model of generation of French prosody", *Speech Communication*, **8**, 137-146.

Bailly, G. ,Laboissière, R. & Schwartz, J-L (1991) "Formant trajectories as audible gestures: an alternative for speech synthesis", *Journal of Phonetics*, **19**, 9-23.

Barbosa, P. (1991) "Génération automatique de la prosodie du français", *Rapport DEA*, ENSERG/ ICP, Grenoble.

Bartkova, K. & Sorin, C. (1987) "A model of segmental duration for speech synthesis in French", *Speech Communication*, **6**, 245-260.

Campbell, W.N. (à paraître) "Syllable-based segmental duration" , in *Talking machines : theories, models and designs* , (Bailly,G. & Benoît, C., Eds).

Di Cristo, A. (1985) *De la microprosodie à l'intonosyntaxe*, Université de Provence.

Emerard, F. (1977) *Synthèse par diphones et traitement de la prosodie, Thèse 3e cycle*, Grenoble, France.

Fraisse, P. (1974) *La psychologie du rythme*, PUF, Paris.

Jordan, M. (1986) "Serial order: a parallel distributed processing approach", *Technical*

Report ICS-8604. La Jolla: University of California - Institute for Cognitive Science.

Lehiste, I. (1977) "Isochrony reconsidered", *Journal of Phonetics*, **5**, 253-263.

Marcus, S. M. (1976) *Perceptual centres, Thèse de doctorat non publiée*, Cambridge University.

O'Shaughnessy, D. (1981) "A study of French vowel and consonant durations", *Journal of Phonetics*, **9**, 385-406.

Pike, K.L. (1945) *The intonation of American English*. Ann Arbor: University of Michigan Press.

Pompino-Marschall, B. (1989) "On the psychoacoustic nature of the P-center phenomenon", *Journal of Phonetics* , **17**, 175-192.

Sagisaka, Y. (1990) "On the prediction of global Fo shapes for Japanese text-to-speech", *Int. Conf. on Acoust. Speech & Sig. Proc.*,**1** , 325-328.

Todd, P.(1989) "A connectionist approach to algorithmic composition", *Computer Music Journal*, **13**, **4**,27-43.

Traber, C. (1990) "Fo generation with a database of natural Fo patterns and with a neural network", in *Talking machines : theories, models and designs* , (Bailly,G. & Benoît, C., Eds).

Turvey, M.T., Schmidt, R.C. & Rosenblum, L. (1990) "Clock and motor components in absolute coordination of rhythmic movements", *Haskins Laboratories Status Report on Speech Research*, 231-242.

van Santen, J.P.H. & Olive, J.P. (1990) "The analysis of contextual effects on segmental durations", *Computer, Speech & Language*, **4**, 359-390.

ANNEXE

TABLE 3. Moyennes et écarts-type en ms des durées phonémiques obtenus.

Phon	Moy	Ec-Type	Phon	Moy	Ec-Type	Phon	Moy	Ec-Type
a	114	40	ɑ̃	137	44	f	143	37
ɛ	109	33	ɛ̃,œ̃	139	56	s	166	48
e	121	51	ɔ̃	139	46	ʃ	142	27
i	101	37	p	123	42	v	105	30
œ	106	39	t	127	39	z	115	25
ø	117	32	k	116	42	ʒ	104	26
y	103	30	b	91	22	m	104	24
ɔ	110	33	d	92	25	n	101	24
o	126	36	g	87	24	ʀ	94	39
u	108	32	j	93	29	l	93	27
w	94	25	ɥ	91	20	ɲ	112	28

19es J.E.P. - Bruxelles, 19 au 22 mai 1992.

Un système de dialogue oral pour une application de réservation téléphonique de billets d'avion

F. CHARPENTIER, F. GAVIGNET, K. CHOUKRI, F. ANDRY,
E. BILANGE, J.-Y. MAGADUR

CAP GEMINI INNOVATION
118, rue de Tocqueville, 75017 Paris, FRANCE.

Résumé

Cet article décrit l'architecture du système de dialogue oral développé pour une application pilote de service de réservation téléphonique. Les premiers résultats indiquent que les taux de reconnaissance et de compréhension atteignent un niveau acceptable, à condition d'appliquer systématiquement une stratégie de prédiction des contraintes lexicales, syntaxiques et sémantiques à partir de la connaissance de l'état courant du dialogue.

1 Introduction

Cet article présente le système de dialogue oral développé pour le français dans le cadre du projet Esprit SUNDIAL[1][1]. L'une des originalités de ce projet par rapport aux autres travaux sur le dialogue oral [2][3][4][5] est de prendre en compte les limitations imposées par le réseau téléphonique, dans l'optique de déboucher sur des techniques utilisables pour des services vocaux réels. Il est probable que l'intégration de la reconnaissance vocale dans ces services accompagnera l'essor général de ce secteur, dû au lancement par FRANCE TELECOM de la taxation kiosque sans limitation de durée.

L'application choisie est la réservation de billets d'avion. Les performances souhaitables du système, en terme de couverture linguistique et d'éventail des phénomènes de dialogue traités, ont été identifiées grâce à une expérience originale de simulation selon la méthode du Magicien d'Oz (*Wizard of Oz*) qui a permis de constituer un corpus de 300 dialogues mettant en situation des utilisateurs potentiels [6].

Le but de cet article est de décrire l'architecture et le fonctionnement du prototype mis en place à partir de l'analyse de ce corpus de dialogues et de présenter les premiers résultats sur son évaluation. Le prototype, développé pour le français par CAP GEMINI INNOVATION, est aussi développé pour l'anglais par LOGICA selon une architecture quasiment identique.

Le noyau du système est un dialogueur, dont la tâche est de gérer les deux canaux d'échange avec l'utilisateur (compréhension des énoncés de l'utilisateur et réponse vocale), ainsi que l'accès à la base de données de vols. Il permet de traiter les phénomènes de dialogues courants (contestation, confirmation, répétition) dans un mode d'interaction ouvert, tout en passant dans un mode directif pour la gestion des erreurs éventuelles de reconnaissance vocale. Le dialogueur reçoit du canal de compréhension une représentation sémantique des énoncés de l'utilisateur, puis il retourne simultanément la représentation du message à générer vers le canal de réponse vocale et une prédiction du prochain énoncé de l'utilisateur vers le canal de compréhension.

Le canal de compréhension comprend un étage de reconnaissance vocale basé sur une modélisation HMM (*Hidden Markov Models*) du vocabulaire de l'application et un étage d'analyse linguistique basé

[1] Ce projet est partiellement financé par le programme ESPRIT de la communauté européenne. Les partenaires du projet sont CAP GEMINI INNOVATION, CNET, CSELT, DAIMLER-BENZ, UNIVERSITÉ D'ERLANGEN, INFOVOX, IRISA, LOGICA, POLITECHNICO DI TORINO, SARITEL, SIEMENS, UNIVERSITÉ DE SURREY

sur une grammaire d'unification catégorielle, qui effectue de façon conjointe l'analyse syntaxique et sémantique.

Le canal de réponse vocale est constitué d'un générateur de messages qui transforme la représentation sémantique interne au dialogueur en un texte orthographique et d'un système de synthèse à partir du texte.

2 La reconnaissance vocale

Le prototype utilise le système de reconnaissance de la parole continue développé par LOGICA [7], que nous avons adapté au français. L'algorithme de reconnaissance est basé sur une modélisation HMM des phonèmes du français, indépendante du contexte phonétique. Ces modèles ont été obtenus par apprentissage sur une base de données de 99 locuteurs enregistrés sur le réseau téléphonique commuté (RTC), chacun ayant prononcé 150 phrases (phrases phonétiquement équilibrées et phrases liées à l'application aérienne) et 180 mots isolés (dont les chiffres, les jours, les mois, et les lettres). La base de locuteurs comporte deux tiers d'hommes et un tiers de femmes.

L'algorithme réalise directement la reconnaissance de mots, sans passer par un intermédiaire phonétique. Les modèles des mots sont formés par concaténation des modèles de phonèmes. Le lexique actuel comprend plus de 300 mots, plus la liste variable des paramètres de l'application (destinations aériennes, etc...). L'introduction d'hypermots obtenus par concaténation de plusieurs mots en un seul modèle HMM, de façon indépendante des niveaux d'analyse linguistique, permet d'améliorer les taux de reconnaissance pour les expressions figées (*je voudrais*, *tout à fait*, etc.).

L'algorithme fonctionne de façon synchrone des trames acoustiques, en utilisant une recherche en faisceau pour limiter le nombre de mots simultanément actifs. Lorsqu'un mot est complètement reconnu, il est enregistré dans un *graphe lexical*, muni des trames de début et de fin, et de ses mots sucesseurs possibles, qui dépendent du modèle de langage utilisé. L'algorithme redémarre les mots successeurs à la trame suivante, en parallèle avec les mots actifs non terminés.

Le modèle de langage utilisé pour limiter la perplexité due à ce redémarrage continuel des reconnaissances lexicales est un ensemble de *grammaires de paires de mots* (listes de séquences de deux mots permises), correspondant aux différentes prédictions émises par le dialogueur. Ces grammaires constituent les *prédictions statiques* du dialogue, et elles sont cruciales pour assurer un taux de reconnaissance suffisant.

Pour les construire, nous avons identifié 16 états caractéristique du dialogue à partir du corpus de dialogues issu de la simulation [6]. Les énoncés de l'utilisateur pour chacun de ces états ont été modélisés sous forme de grammaires hors-contexte. Puis tous les énoncés possibles ont été générés à partir de ces grammaires, dans le but de construire les sous-lexiques et les grammaires de paires de mots associées aux états de dialogue. Cette méthode de précompilation des modèles de langage est analogue à celle proposée dans [3].

L'importance de prédictions statiques est illustré dans le tableau de résultats suivant, pour un corpus de test de 72 phrases prononcées par un locuteur masculin. On tient compte également de la qualité du canal acoustique (PABX ou RTC). On donne ici les taux de reconnaissance de phrases, le taux de reconnaissance lexicale (*word accuracy*) étant précisé entre parenthèses.

	PABX	RTC
Sans prédiction	14 (36)	5 (25)
Avec prédiction	57 (86)	49 (79)

3 L'analyseur linguistique

Le module d'analyse linguistique reçoit en entrée le graphe lexical construit par le module de reconnaissance et produit deux choses : l'arbre syntaxique et la représentation sémantique de la meilleure hypothèse exprimée.

3.1 Le formalisme grammatical.

Notre formalisme grammatical est dérivé des **UCG** (Grammaires Catégorielles d'Unification) [8]. Dans ce formalisme, toutes les informations morphologiques, syntaxiques et sémantiques sont contenues dans les entrées lexicales sous forme d'attributs valués. La grammaire est quant à elle réduite à quelques règles de combinaison basées sur l'unification. Cette capacité d'intégrer différents niveaux linguistiques dans une structure de données unique, est très appréciable pour le traitement de la parole. En effet, le rôle de certaines connaissances spécifiques peut être ajusté pour couvrir l'échec de connaissances d'un autre niveau.

Ainsi les déterminants peuvent être optionnels pour les noms puisque ceux-ci sont habituellement courts et donc difficiles à détecter par le module de reconnaissance. A l'inverse, les contraintes sémantiques entre le nom et l'adjectif peuvent être renforcées. D'une façon générale, nous réduisons

la possibilité de sur-générer des hypothèses lors de l'analyse.

Le formalisme UCG permet de représenter efficacement les constituants optionnels d'un syntagme et ceux qui peuvent apparaître en ordre variable. Cette propriété est particulièrement intéressante pour la compréhension de la parole, où l'ordre des constituants est très libre. L'optionalité des arguments est incluse dans un trait particulier. Une facette additionnelle indique si l'argument en question doit être adjacent ou non.

Ainsi pour le verbe "*rentrer*", nous sommes en mesure de construire un certain nombre de combinaisons comme pour :

- *rentrer avant vingt heures*
- *rentrer de Stuttgart avant vingt heures*
- *rentrer le 5 mars de Stuttgart*
- *rentrer de Stuttgart le 5 mars*

Le générateur de lexique, développé spécifiquement pour le projet [9], permet de construire le lexique UCG. Pour ce faire, les connaissances linguistiques du français sont combinées aux connaissances du domaine (un réseau sémantique muni de liens d'héritage et de rôles typés). Le ratio entre le nombre d'items lexicaux obtenus pour la grammaire UCG et le nombre de formes fléchies différentes est de 1,5.

3.2 Analyse grammaticale

Afin d'optimiser le temps de traitement, nous avons divisé notre mécanisme d'analyse en *deux phases* [10].

Dans une première phase, nous utilisons un ensemble réduit de contraintes syntaxiques et sémantiques pour déterminer une liste de phrases grammaticalement acceptables, ordonnée par scores décroissants. Cette phase, dite *phase d'acceptation*, a été optimisée par l'utilisation de vecteurs de bits. En effet, ce type de représentation est extrêmement compact. De plus, l'unification dans un langage évolué de type C se réduit, pour ces vecteurs, à des opérations booléennes au niveau de la machine.

Dans la seconde phase, *phase de compréhension*, une représentation sémantique de la meilleure acceptée est réalisée. Ceci permet de retarder les opérations complexes et coûteuses de construction de structure jusqu'à ce qu'une solution soit trouvée (voir figure 1).

Les deux phases sont basées sur la même stratégie: analyse gauche-droite ascendante. Un diagramme actif (*chart*), dont les arcs sont étiquetés avec les scores acoustiques, est construit. La recherche de la solution au cours de l'analyse s'effectue par l'intermédiaire d'un faisceau d'arcs. Bien que

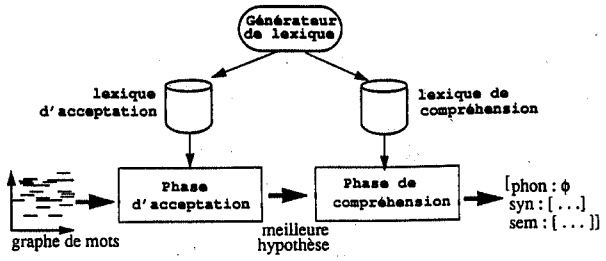


Figure 1: Composants de l'analyseur de SUNDIAL

ce type de recherche soit non admissible, des expérimentations ont permis de déterminer la taille optimale du faisceau à 25 arcs environ. Le calcul du score des hypothèses est obtenu pendant la première phase de l'analyse, juste après l'unification.

3.3 Interaction avec le dialogueur

La communication avec le dialogueur a lieu dans les deux sens : d'une part la transmission de la représentation syntaxique et sémantique de l'interprétation de la phrase émise par l'utilisateur, et d'autre part le retour d'informations du niveau dialogue vers l'analyseur sous la forme de prédictions.

Le Langage Sémantique d'Interface (**SIL**) [11] permet de construire des expressions qui décrivent les objets du domaine sous la forme de structures récursives de couples attribut-valeur. Les expérimentations ont montré que SIL répond bien aux besoins de représentation. Toutefois, les aspects temporels (date et heures), ainsi que la représentation des nombres qui utilisent plus d'une trentaine de concepts ne nous donnent pas encore complète satisfaction et demandent à être affinés.

A chaque intervention du système, le dialogueur est en mesure de prédire des informations sur le contenu des interventions de l'utilisateur un *contenu intentionnel* composé d'actes de dialogue, et un *contenu propositionnel* de nature sémantique sur lequel porte les actes. Deux types de mécanismes sont mis en oeuvre pour l'exploitation de ces prédictions [12].

La détermination de l'état du dialogue courant permet de sélectionner la bonne prédiction statique, sous la forme des sous-lexiques et des grammaires de paires de mots discutées plus haut.

D'autre part, ces prédictions contribuent à améliorer les performances de compréhension en permettant d'extraire les phrases contextuellement pertinentes de la liste de phrases produites par la phase d'acceptation de l'analyseur.

Des contraintes sémantiques sont élaborées à partir du focus et de ses glissements possibles, construits de façon dynamique. Des contraintes dialogiques liées à des marqueurs caractéristiques sur les

formes de surface viennent compléter ces restrictions sémantiques.

Les résultats suivants permettent d'apprécier le gain de compréhension apporté par l'étage d'analyse linguistique et par les prédictions dynamiques effectuées en fin de phase d'acceptation. Le test de compréhension porte sur 72 phrases prononcées par un locuteur masculin pour une communication interne via le standard téléphonique (qualité PABX):

Etape	Taux de reconnaissance Phrases (Mots)
Reconnaissance et prédictions statiques	57 (86)
Analyse linguistique	65 (80)
Prédictions dynamiques	74 (83)

4 Le dialogueur

4.1 Architecture interne

Le dialogueur est constitué de cinq modules. Ces modules communiquent les uns avec les autres afin de décider des actes de dialogue qui doivent être produits par le système à chaque tour de parole. Chaque module a un domaine d'expertise bien précis :

- Le module de tâche est seul à avoir accès à la base de données . Il cherche à obtenir une solution, ou un ensemble de solutions, correspondant à la requête de l'utilisateur. Dans ce but, il interroge le module de dialogue pour acquérir les paramètres nécessaires à un accès à la base de données. Une fois que la solution a été obtenue et validée par l'utilisateur, il met à jour la base de données. Le module de tâche peut fonctionner avec plusieurs applications différentes, sans qu'il soit pour cela nécessaire de modifier les autres modules du dialogueur. Ainsi l'application retenue par les partenaires allemands concerne les horaires de trains et celle des partenaires anglais les horaires d'avions (sans réservations).

- Le module de dialogue a pour rôle de maintenir la cohérence du dialogue à un niveau global. Il décide, à partir des informations reçues du monde extérieur, de la liste des actes de dialogue qui doivent être produits par le système à chaque tour de parole. Les deux sources d'information extérieures essentielles pour le module de dialogue sont les messages envoyés par le module de tâche et les expressions SIL représentant les énoncés de l'utilisateur, envoyées par l'interface linguistique. C'est lui qui assure l'interprétation des actes de dialogue produits par l'utilisateur, ainsi que le calcul des prédictions sur les actes futurs de l'utilisateur. Le modèle théorique de dialogue qui sert de base à ce module, ainsi que les détails de son implémentation, sont présentés dans [13].

- Le module de croyance est une base de données qui maintient une représentation des objets et des relations du domaine. Il permet aux autres modules d'accéder à tout instant à l'état courant des connaissances mutuelles du système et de l'utilisateur à propos des objets du domaine.

- L'interface linguistique est un module d'interfaçage avec l'étage d'analyse linguistique. Il reçoit l'expression SIL représentant l'énoncé de l'utilisateur et, après avoir mis à jour l'historique linguistique, la transmet au module de dialogue.

- Le planificateur de messages construit la représentation sémantique du prochain énoncé du système à partir de la liste des actes de dialogue à produire et de leur contenu propositionnel, listé envoyée par le module de dialogue.

Schématiquement, on peut décrire le fonctionnement du dialogueur de la façon suivante.

L'expression SIL produite par l'analyseur est envoyée à l'interface linguistique qui, après avoir mis à jour l'historique linguistique, la transmet au module de dialogue. Celui-ci, sur la base des prédictions produites au tour de parole précédent, et avec l'aide du module de croyances, interprète l'entrée de l'utilisateur, c'est à dire qu'il lui associe un acte de dialogue et qu'il place celui-ci dans la structure de dialogue. Les informations acquises (par exemple la valeur d'un paramètre de la tâche) sont transmises au module de tâche qui en vérifie la cohérence, et émet éventuellement de nouvelles requêtes. Ces requêtes conduisent le module de dialogue à la production de nouveaux actes de dialogue, qui sont envoyés, avec leur contenu propositionnel, au module de planification de messages. Celui-ci génère la représentation sémantique de l'intervention du système, et l'envoie à l'étage de synthèse vocale. En même temps qu'il produit la liste des actes à générer, le module de dialogue utilise la grammaire de dialogue pour faire des prédictions sur les actes à venir de l'utilisateur. Ces prédictions, qui consistent en une liste d'actes prédits avec un contenu propositionnel associé, sont envoyées au module d'analyse linguistique.

4.2 Phénomènes traités

Le dialogueur est actuellement capable de prendre en compte plusieurs phénomènes de dialogue importants, tels que la gestion de contestations de

l'utilisateur, d'échanges évaluateur imbriqués, de réintroductions d'initiatives. Le dialogue suivant est, par exemple, supporté par le système actuel :

S_1 *Bonjour, formulez votre demande*
U_1 *Je voudrais réserver un vol Paris-Londres*
S_2 *Paris-Londres. Quel jour ?*
U_2 *Le 6 juin.*
S_3 *le 6 juin. A quelle heure ?*
U_3 *pas le 6, le 10*
S_4 *le 10 juin ?*
U_4 *oui, le 10 juin.*
S_5 *et à quelle heure voulez-vous partir ?*

Dans cet exemple, le système confirme systématiquement tous les paramètres qu'il acquiert. De plus, si le système doit acquérir une nouvelle information, et s'il ne se trouve pas dans un échange clarificateur, il utilise le même tour de parole pour produire la confirmation et la question (comme dans S_3). Cette stratégie permet de s'assurer continuellement de la fiabilité des informations reçues sans rompre pour autant la fluidité du dialogue. D'autres stratégies de confirmation sont possibles. Ainsi, dans le prototype developpé par les partenaires allemands, on consacre un tour de parole à la confirmation de chaque paramètre. Puis, lorsque le ou les paramètres ont été confirmés, le système reprend initiative.

Le système est aussi doté de capacités de réparation d'erreurs. Lorsqu'il détecte plusieurs échanges évaluateurs imbriqués dans la structure de dialogue (ce qui indique la présence d'un problème de communication), il entre dans un mode de réparation d'erreur. Si le problème porte par exemple sur un nom de ville, le système demande à l'utilisateur d'épeler le nom en question.

5 Epellation sur des vocabulaires fermés

De façon générale, une classe importante de paramètres à saisir dans un dialogue oral correspond aux entités désignées par leur nom, comme précisément les villes desservies par des réseaux de transport. Dans le cas du transport aérien, ADP recense quelques 300 villes de destinations, tandis que pour le chemin de fer, la SNCF recence quelques 5000 gares. Or, bien que ces paramètres soient sollicités en mode ouvert dans le dialogue, il concernent un vocabulaire limité, contrainte que l'on peut exploiter pour la récupération d'erreur, en passant en mode d'épellation. Cette possibilité est paradoxale étant donné les mauvais taux de reconnaissance des lettres en isolation (de l'ordre de 60% pour notre système). Mais les contraintes imposées par le lexique sont suffisamment fortes pour permettre un très bon taux de reconnaissance global.

Nous avons expérimenté l'idée sur les destinations d'une grande compagnie aérienne (166 villes), en utilisant un mode simplifié d'épellation, par lettres isolées. L'algorithme utilise comme point de départ la matrice de confusion des lettres épelées, obtenues lors des tests de l'étage de reconnaissance. A la fin de l'épellation, chacune des villes possibles reçoit un score par addition des coefficients de confusion entre ses propres lettres et les lettres reconnues. Ce mécanisme permet alors de retrouver la ville épelée avec une grande fiabilité, même si plusieurs lettres sont mal reconnues. Par exemple, *p.a.r.i.s.*, qui est reconnu comme la séquence *t.h.r.i.f.* sans appliquer les contraintes lexicales, est identifié correctement après reconnaissance globale. Lorsque ce procédé ne fournit pas la bonne réponse en première position, il effectue un ordonnancement des villes suffisamment efficace pour que le second candidat soit le bon. Aussi le dialogue utilise-t-il les résultats de l'épellation dans un mode de confirmation explicite: le système demande à l'utilisateur de valider le premier cándidat, sinon le deuxième,...

Cependant, l'algorithme actuellement utilisé est trop discriminant car il identifie de façon implicite le nombre de lettres de la ville, sans tenir compte des risques d'insertion ou d'élimination de lettres. Un algorithme de programmation dynamique permettrait de traiter le problème avec plus de souplesse. Mais, l'objectif pratique est de permettre l'épellation en mode continu, ce qui nécessite l'identification dans le graphe de lettres des meilleurs chemins acoustiques correspondant aux villes permises.

6 Génération des messages oraux

Pour générer les messages du système, nous avons adopté une approche pragmatique, de génération par schémas de phrases contenant des parties variables instanciées par les paramètres de la tâche. Les messages sont définis de manière déclarative dans une table où les schémas de phrases sont associés à des représentations sémantiques composées des actes de dialogue utilisés par le dialogueur et de couples *attribut = valeur*, représentant à la fois les paramètres associés et des contextes du dialogue. Cette approche permet une grande souplesse pour la mise au point et la modification d'un corpus de messages en cours d'expérimentation.

Ce module de génération fournit la possibilité d'utiliser des formulations différentes en fonction du contexte du dialogue pour exprimer le même contenu sémantique: de choisir, par exemple, entre des mes-

sages de confirmation plus ou moins explicites. De plus, le formalisme de déclaration des messages inclut la notion de paramètre optionnel ce qui permet de générer un grand nombre de messages différents à partir d'une définition compacte des schémas de messages.

Les messages générés sont ensuite vocalisés par le système de synthèse à partir du texte du CNET, fourni dans le cadre de SUNDIAL, utilisant les algorithmes de synthèse PSOLA et les algorithmes linguistiques CNETVOX.

7 Conclusion

Une première version du prototype a été mise en place, qui intègre les composantes vocales du système (reconnaissance et synthèse). Il fonctionne sur le téléphone, en mode parole connectée, et illustre la fonction de saisie de paramètres (heures, dates et villes), ainsi que les fonctions de récupération d'erreurs de reconnaissance du type épellation ou recherche guidée.

La seconde version du système, fonctionnant en mode parole continue, intègre l'analyseur linguistique et le dialogueur.

L'utilisation des prédictions émises par le dialogueur semble indispensable, à cause des performances limitées de la reconnaissance sans modèle de langage. Mais sous cette condition, les premiers résultats obtenus pour le canal de compréhension sont encourageants. En effet, ils indiquent un taux de reconnaissance de phrases acceptable pour la conduite du dialogue (57%), et un gain de compréhension substantiel apporté par les étages linguistiques (74%).

Des résultats sur les performances globales du dialogue oral, non encore disponibles lors de la rédaction de cet article, seront présentés lors de la conférence.

Remerciements

Ce travail est par nature collectif et a bénéficié de la collaboration avec les autres équipes impliquées dans SUNDIAL. Nous remercions plus particulièrement nos partenaires de LOGICA, du CNET et de l'IRISA qui ont contribué directement au prototype français, et nous voudrions remercier tout particulièrement T. Thomas, S. Thornton, A. Cozannet et D. Sadek pour leur aide précieuse.

References

[1] Peckham J., "Speech Understanding and Dialogue over the telephone : an overview of the ESPRIT SUNDIAL project", 2nd European Conference on Speech Communication and Technology, pp. 1469-1472, Genova, 24-26 Sept. 1991.

[2] Matrouf A.K., "Un système de dialogue oral orienté par la tâche", Thèse de l'Université Paris-Sud, Sept. 1990.

[3] Zue, V. et al. (1991) Integration of speech recognition and natural language processing in the MIT Voyager System, *Int. Conf. on ASSP* , 713-717.

[4] Young S.J., Proctor C.E., "The Design and Implementation of Dialogue Control in Voice Operated Database Inquiry Systems", *Computer Speech and Language*, 3, pp. 329-353, 1989.

[5] Guyomard M., Siroux J., Cozannet A., "Le rôle du dialogue pour la reconnaissance de parole. Le cas du système pages jaunes", *XVIII èmes journées d'études sur la parole*, 28-31 Mai 1990, Montréal, Canada.

[6] Andry F., Bilange E., Charpentier F., Choukri K., Ponamalé M., Soudoplatoff S., "Computerised simulation tools for the design of an oral dialogue system", Proc. of ETW Conference, Bruxelles, 1990.

[7] Micca G. & al., "Intermediate report on front end processing techniques", Deliverable WP4-D2, SUNDIAL Project, June 1990.

[8] Zeevat H., Klein E. and Calder J., "An Introduction to Unification Categorial Grammar", in Haddock N., Klein E. and Morill G. (eds.) *Edinburgh Working Papers in Cognitive Science*, V.1 : Categorial Grammar, Unification Grammar and Parsing, 1987.

[9] Andry F., Fraser N., Mcglashan S., Thornton S., Youd N., "Constructing linguistic Knowledge bases for applications : a general-purpose tool", *Computational Linguistics, special issue on Inheritance in Natural Language Processing*, Walter Daelemans and Gerald Gazdar eds., 1991.

[10] Andry F., Thornton S., "A parser for speech lattices using a UCG grammar", 2nd European Conference on Speech Communication and Technology, pp. 219-222, Genova, 24-26 Sept. 1991.

[11] McGlashan S., Andry F. and Niedermair G., *A proposal for SIL*, SUNDIAL technical report, Dec 1990.

[12] Andry F. "L'utilisation des prédictions dans le dialogue oral homme-machine", Thèse de l'université Paris XIII, 1992.

[13] Bilange, E. (1991) Modélisation du dialogue oral finalisé personne-machine par une approche structurelle, théorie et réalisation. Thèse de l'Université de Rennes I, Décembre 1991.

19es J.E.P. - Bruxelles, 19 au 22 mai 1992.

ICPplan : DIALOGUE MULTIMODAL POUR LA CONCEPTION DE PLANS ARCHITECTURAUX

M.L. BOURGUET

ICP / INPG, CNRS URA n° 368
UNIVERSITE STENDHAL,
46 AV F. VIALLET 38031 GRENOBLE CEDEX FRANCE

Résumé

L'aspect multimodal d'un système de dialogue homme-machine soulève de nombreux problèmes depuis la prise en compte des événements circulant sur les médias jusqu'à leur interprétation au cœur de l'application support. Le choix du type de multimodalité (exclusif, concurrent, alterné, composé) influe sur le résultat de l'interprétation. L'information gestuelle attachée à un acte de désignation n'a pas le même degré d'ambiguïté que celle d'un acte oral. Ces deux modes, lorsqu'ils coopèrent dans l'expression d'un message à travers des déictiques, sont étroitement imbriqués. Nous montrons que le niveau morphologique est un bon niveau pour les fusionner. Le deuxième but de l'article est de montrer qu'une description de la tâche en termes hiérarchiques d'actions-actants, de scripts, de scénarios et de plans offre une solution aux problèmes de coréférences, de conflits et de redondance entre les modes A titre d'illustration, nous présentons l'application ICPplan.

1. INTRODUCTION

La communication homme-machine est appelée multimodale lorsqu'elle met en jeu plusieurs canaux sensori-moteurs de l'homme — vision, parole (entendue et produite), geste (mouvement, désignation, écriture, dessin), etc. — en lui autorisant plusieurs modes d'expression simultanés et complémentaires (Taylor & al, 1989). L'information parvient à la machine par plusieurs canaux (médias). Ces nouvelles performances de la machine semblent séduisantes à priori; on escompte ainsi améliorer l'efficacité de l'interaction (entrées de plusieurs commandes simultanément), sa fiabilité (utilisation de la redondance), sa souplesse (choix des modes de communication les mieux adaptés à la tâche), en un mot son ergonomie (Falzon, 1990). Mais à y regarder de plus près est-on sûr de réduire la charge cognitive de l'utilisateur ? Ne risque-t-on pas d'embrouiller ses schémas de planification et ses modèles de représentation ? Il faut remarquer avant tout que ces nouvelles interfaces nous présentent des mondes inhabituels peuplés d'objets qui peuvent réagir à plusieurs commandes simultanément, par exemple je peux tracer à main levée une courbe sur l'écran tout en prononçant les mots "rouge" puis "vert" pour obtenir une seule courbe en deux couleurs... ce qui n'est guère possible dans le monde réel du crayon et de la feuille de papier. On voit donc par cet exemple simple qu'une réflexion est à mener sur la pertinence de la multimodalité en communication homme-machine.

Nous allons examiner ci-après comment les divers types de multimodalités peuvent intervenir dans une interface homme-machine et proposer des solutions pour la gestion des événements et l'interprétation des informations provenant de chaque médium, car les solutions choisies inter-agissent à leur tour sur les stratégies profondes mises en œuvre dans le dialogue.

Tout d'abord, il est important de distinguer clairement les termes multimédia et multimodal (médium et mode) puisqu'ils s'opposent sur les plans de la forme et du contenu (Coutaz & Caelen, 1990):

- un *médium* est un support, il se définit par le type de capteur ou d'effecteur qu'il met en oeuvre. Un système multimédia ne traite des informations qu'à travers leur forme, il les manipule mais ne les comprend pas.

- un *mode* d'expression sous-tend l'idée de langage qui véhicule des signes. Ces signes présentent une forme et un contenu. En ce sens un système multimodal est

certes multimédia mais doit offrir des capacités de compréhension (traiter le contenu) de plusieurs langages (parole, écrit, geste, etc.), chacun d'eux se définissant par un lexique, une syntaxe, une sémantique, etc. Symétriquement, un système multimodal doit être capable également de présenter sous divers points de vue, des informations visuelles, auditives, etc. selon la tâche à accomplir et la charge cognitive supposée de l'utilisateur.

Il est possible de distinguer quatre types de systèmes multimodaux (Workshop IHM'91), selon que les différents médias sont accessibles de façon séquentielle ou parallèle, et selon qu'un même message est distribué ou non sur plusieurs modes ce qui nécessite (ou non) la prise en compte de tous les modes pour interpréter ou générer un message (figure 1). Il est évident que nous ne parlons ici que des entrées du système puisqu'il est courant maintenant de ventiler les sorties sur différents périphériques (son, image, etc.).

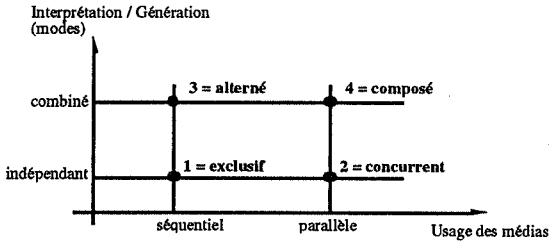


Figure 1: Les types de systèmes multimodaux

1 • *Multimodal exclusif* : deux médias ne peuvent pas être utilisés en même temps et les informations véhiculées par deux modes différents restent indépendantes, c'est par exemple le cas des sytèmes informatiques classiques, où il est possible d'entrer une commande au clavier ou de sélectionner une action dans un menu déroulant,

2 • *Multimodal concurrent* : l'usage des médias peut se faire de façon parallèle, mais les informations circulant sur ces médias sont indépendantes. Il peut y avoir redondance, par ex. lorsqu'on affiche un texte à l'écran et qu'on le synthétise aussi par la parole, ou conflit si deux commandes contradictoires surviennent en même temps. Il est difficile d'imaginer que l'on pourrait avoir une commande comme "mets çà là", car le déictique "çà" réfère habituellement à un geste et donc ici à une information d'une autre canal. Dans ce type de multimodalité un geste survenant pendant le temps de parole serait privé de toute coréférence à la parole et il en résulterait une désynchronisation désagréable pour l'utilisateur,

3 • *Multimodal alterné* : l'usage des médias est séquentiel, et le traitement des informations peut combiner les différents modes. Par ex. je dis "mets çà là" et je désigne gestuellement l'objet "çà" et le lieu "là", après la fin de la phrase. Dans un tel système un côté artificiel dans l'usage des modes subsiste au niveau de la synchronie des différents actes élémentaires, bien qu'ils puissent être entrelacés pour la constitution d'un message. Malgré les limitations apparentes de ce type de système, cet usage alterné des médias réduit l'ambiguïté d'interpénétration des informations,

4 • *Multimodal composé ou synergique* : l'usage des médias est parallèle et les traitements sont combinés. On atteint ici le niveau le plus intéressant pour une communication naturelle puisque l'acte verbal "mets çà la" peut être réalisé en synergie avec les actes gestuels afférants. Mais on atteint également un niveau de complexité et d'ambiguité qui n'existaient pas dans les autres types: en effet si je dis "déplace les objets" et que je sélectionne des objets par la souris, on ne sait pas si ces deux actions sont liées ou non car il y a plusieurs interprétations possibles du fait que deux actions parallèles sont autorisées.

2. EVENEMENTS ET INFORMATION

Le problème de l'interprétation d'un message multimodal composé résulte du fait que des informations partielles circulent sur des canaux différents et sont exprimées dans des langages différents (Caelen, 1991) — par exemple parole et geste auxquels nous nous restreindrons dans la suite de cet article. Nous distinguerons tout d'abord les notions d'événement et d'information :

• un événement, pris au sens du génie logiciel, sous-tend une notion d'interruption, de canal de communication (médium) et de date. Un événement est donc un signal, ce n'est pas un signe au sens linguistique du terme. Nous définissons alors:

dK souris enfoncée, fK souris relâchée, K mouvement souris,
dP début parole, fP fin parole,
dC début message clavier, fC fin message clavier,
dS touche "shift" enfoncée, fS touche "shift" relâchée,
comme noms symboliques d'événements attachés à une date et une provenance,

• les informations qui circulent à travers l'interface sont structurées et ont un sens dans un code linguistique donné (langue parlée, langage gestuel). Ces informations sont donc les éléments

d'un langage, qui présente les couches de structuration habituelles : morphologie, syntaxe, sémantique et pragmatique. A la parole et au geste sont associés respectivement un langage Lp et un langage Lg.
Dans la plupart des systèmes multimodaux actuels (MMI^2 par exemple) (Falzon, 1991) la tendance est de traiter chacun des langages pris séparément puis d'effectuer une interprétation coréférentielle finale. Nous pensons au contraire que les informations doivent être fusionnées au plus bas niveau, c'est-à-dire au niveau morphologique pour éviter les lourdeurs de la modalité alternée (voir ci-dessus). Cela consiste donc :

— pour un geste

à associer des événements dK, K et fK formant une suite morphologiquement cohérente et constituant un "morphème" asocié à un item de l'interface (point de coordonnée, zone écran, etc.). Ces suites peuvent se mettre sous forme de structures parenthétiques complètes (fig. 2),

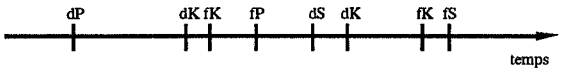


Figure 2: Granularité des événements pour une interface parole/souris. Les suites cohérentes sont : (dP (dK, fK) fP) (dS (dK fK) fS). Pour la suite (dK, fK) l'analyse morphologique consiste à rechercher le point de coordonnées concerné par le geste souris et pour (dS (dK, fK) fS) l'ensemble des points sélectionnés dans le déplacement de la souris.

— pour la parole

à analyser entre les événements dP et fP tous les mots (morphèmes) de la commande. Cette analyse est une simple analyse morphologique donnant éventuellement plusieurs solutions concurrentes, par exemple "Déplace les cercles" donne :
Déplacer (Verbe (impératif, 1ère pers.+sing.)) le (article (déf.+plur.) OU pronom) cercle (Substantif (masc.+plur.))

La "fusion" des informations se fait alors sur ces données, en associant aux déictiques de l'énoncé oral, les suites gestuelles morphologiquement compatibles. Une analyse linguistique interprétative peut alors véritablement commencer sur ces données unifiées (syntaxe, sémantique et pragmatique).

Exemples d'interprétation

Dans l'exemple de la figure 3, l'usage des médias est parallèle. Deux types de multimodalité (concurrent ou composé) peuvent être mis en oeuvre:

— si le système est du type multimodal concurrent, il distingue deux commandes indépendantes et traite séparément les informations qu'il reçoit simultanément sur deux médias distincts. Selon que des objets "triangle" ont été désignés ou non dans l'énoncé précédent, le système interprètera le mot "les" comme une anaphore faisant référence à ces triangles, ou décidera de pivoter tous les triangles visibles sur l'écran.

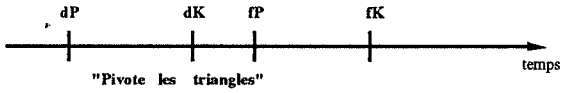


Figure 3: Séquence d'événements pour une interface multimodale concurrente ou composée : l'utilisateur sélectionne en les entourant plusieurs objets visibles (suite dK, fK) et prononce en même temps le message "pivote les triangles" (suite dP, fP).

— si par contre le système est du type multimodal composé, la fusion des informations provenant des différents modes doit être faite au plus bas niveau de la manière suivante

1) la suite (dK, fK) produit un ensemble de points (x, y) pour lesquels l'analyse morphologique consiste à retrouver les objets de l'application qui ont été entourés par le mouvement de la souris,

2) l'analyse morphologique de l'énoncé oral donne:
Pivoter (Verbe (impératif, 1ère pers.+sing.)) les (article (déf.+plur.) OU pronom) triangle (Substantif (fémin.+plur.))
Plusieurs interprétations sont alors possibles. Cela peut signifier :
1) pivoter les objets sélectionnés du type triangle (les = déictique),
2) pivoter les triangles (les = anaphore) dont il vient d'être question précédemment et sélectionner d'autres objets pour la commande suivante,
3) pivoter tous les triangles (les = défini, pluriel) et sélectionner de nouveaux objets.

Les interprétations 2 et 3 sont identiques à celles que l'on obtient en multimodal concurrent. On considère dans ce cas que l'utilisateur désire mener deux actions en parallèle.
L'interprétation 1 est obtenue par la fusion, au niveau morphologique, du déictique "les" avec l'ensemble des objets sélectionnés par le médium souris. Il reste ensuite, grâce à une analyse linguistique de plus haut niveau (sémantique et pragmatique) à résoudre les éventuels conflits entre informations ou actions contradictoires.

Les difficultés d'interprétation dans ce type de système montrent qu'il faut s'attendre à des attitudes différentes et des plans d'action différents selon les utilisateurs (Valot, 1991). Nous suggérons de résoudre ces problèmes en nous fondant sur un calcul intentionnel de manière à replacer l'action dans son contexte de tâche. Toute action de l'utilisateur est interprétée comme étant un élément d'un scénario ou d'un plan inféré lors des actions précédentes.

3. REPRESENTATION DE L'INFORMATION ET DESCRIPTION DE LA TACHE

La tâche est décrite en termes hiérarchiques de plans, sous-plans, scénarios, scripts et actions. Nous définissons l'action comme une structure informative spécialisée renvoyant à des actants (au sens de la grammaire de cas (Fillmore, 1968) mais définis stricto sensus pour notre application). Ces actants sont identifiés par un code et possèdent un contenu exprimé dans un formalisme connu et interprétable par la machine. Ils participent à la définition des scripts dont ils constituent les diverses facettes.
Les *scripts* sont des schémas dont les attributs décrivent l'ensemble des informations nécessaires à l'exécution d'une tâche élémentaire (ouvrir un fichier, déplacer un objet...) Ces attributs possèdent généralement plusieurs facettes exprimées dans le même formalisme que les "actants".
Les *scénarios* sont définis comme des chemins privilégiés ou des séquences probables de scripts. Ils sont décrits à l'aide de scores que l'on attribue aux différentes successions possibles de scripts au cours du dialogue. De la même façon, les *plans* et les *sous-plans* sont définis en termes de probabilités attachées aux diverses séquences de scénarios.

Les exemples d'interprétation précédents nous ont montré que des ambiguités fortes subsistent si l'on ne tient pas compte du contexte de la tâche. Pour cela le décodage des informations doit faire intervenir les différents modèles de langage (gestuel, oral ...) et les connaissances pragmatiques en tentant, par une double stratégie, ascendante et descendante, de reconnaitre les actions et les scripts puis de guider localement l'interprétation en suivant les prédictions du plan général.

Les deux stratégies sont :

1. Une *stratégie ascendante* qui consiste à retrouver une séquence d'actions et d'actants, puis à déduire de cette séquence un script et un scénario pour inférer un plan et déterminer les intentions de l'utilisateur.

2. Une *stratégie descendante* lorsque la connaissance des intentions de l'utilisateur permet d'activer à l'avance des scénarios et des scripts. Les événements sont alors gérés de façon à retrouver dans les commandes les informations ou actants qui complètent au mieux ces scripts. Cette stratégie de gestion des événements permet le cas échéant de choisir judicieusement parmi les hypothèses issues du système de reconnaissance de la parole, la phrase qui s'inscrit le mieux dans le contexte du dialogue. On peut envisager une coopération étroite entre le dialogueur et le reconnaisseur, ce premier étant capable de corriger ou de contraindre ce dernier (au moyen de restrictions lexicales par exemple).

Les deux stratégies de gestion des événements ne s'excluent pas mutuellement. Elles peuvent au contraire intervenir alternativement selon l'état et les besoins du dialogue.

4. UN EXEMPLE : ICPplan

ICPplan est un logiciel d'aide à la conception de plans architecturaux, qui met en oeuvre une interface personne-machine multimodale du type "composé" et un dialogue coopératif. Le système offre des outils graphiques pour la conception de dessins (Coutaz, 1990) et un savoir en matière d'architecture pour vérifier la cohérence des plans, inférer les intentions du concepteur et lui fournir de l'aide. L'application est du type "dirigée par la tâche". L'interface multimodale propose pour l'interaction personne-machine : le geste, la parole et la langue naturelle écrite en entrée, la vision et la parole en sortie.
Les plans de l'utilisateur pourront être du type : construire un batiment, modifier l'ameublement d'une pièce, etc. Ces plans sont définis comme une succession de scénarios, par exemple : construire

des murs, créer et positionner des objets dans une pièce, etc. Les scénarios sont eux-mêmes définis à partir de scripts dont la figure 4 propose un exemple (script de l'action "créer").

action	A1 {CREER, DESSINER, METTRE, FAIRE} / hist.action {CREER}
quoi	A5 .nom (générique) /hist.quoi
taille	A5 .taille (% quoi %) / A3 (% quoi %) / hist.taille (% quoi %)
nombre	A5 .nombre /défaut {UN}
où	A2 / A3 / A6 / A6 A5 / A6 A4 / A6 A2

avec: / = ou, hist = historique, () = condition, % = vérifier cohérence.

Figure 4: Script de l'action "créer".

Dans le formalisme des scripts, on voit apparaitre à travers leur code (A1, A5, etc.) les actions et actants définis précédemment. Nous avons distingué dans ICPplan six types d'actants (figure 5) pour représenter tous les types d'information nécessaires à la définition compète d'une action. Il s'agit de:
- A2: *lieu ponctuel* (donné par une suite (dK, fK) lorsque l'utilisateur a enfoncé puis relaché la souris sur un point de l'écran),
- A3: *espace écran* (donné par une suite (dK, K, fK) lorsque l'utilisateur a enfoncé, déplacé puis relaché la souris, sélectionnant ainsi un ensemble de points c'est à dire une zone de l'écran),
- A4: *désignation objet* (l'utilisateur à "pointé" un objet de l'application avec la souris),
- A5: *description objet* (l'utilisateur a décrit oralement un objet de l'application, objet déjà instancié (visible sur l'écran) ou non (qui n'existe pas encore)),
- A6: *description lieu* (l'utilisateur a décrit oralement une zone de l'écran, généralement relativement à un objet ou à un espace déjà circonscrit, par exemple : en haut à droite (sous entendu dans la pièce où le curseur se déplace à cet instant) ou à coté de (suivi de la description d'un objet réel)),
- A7: *modulateur* (toute information complémentaire pouvant modifier ou préciser l'action, ex: "je voudrais" (intention), "un peu" (quantité, manière), "encore" (réitération), etc.).

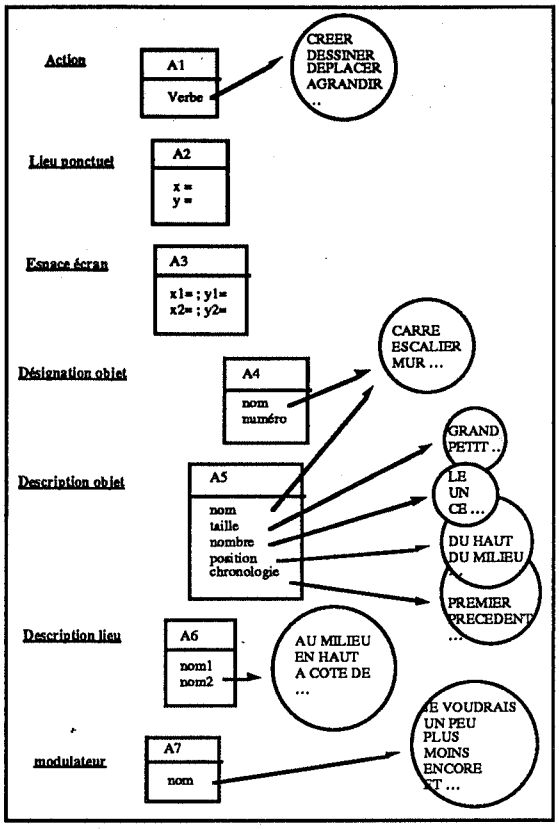
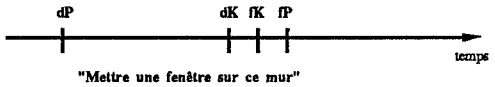


Figure 5: Actions et actants dans ICPplan.

Exemple d'interprétation dans ICPplan



	Mettre	une	fenêtre	sur	ce	mur	dK, fK
Analyse morphologique	Verbe infinitif	article indéfini féminin singulier	substantif féminin singulier	adverbe lieu	déictique masculin singulier	substantif masculin singulier	mur n° 3
Analyse syntaxique	V	COD		CL			
Analyse sémantique (actants)	A1	A5		A6	A4		

Figure 6: Séquence d'événements et interprétation dans ICPplan: l'utilisateur prononce le message "Mettre une fenêtre sur ce mur" (suite dP, fP) et sélectionne en même temps un objet sur l'écran (suite dK, fK).

L'analyse morphologique des commandes permet la fusion des informations multimodales c'est à dire l'association du déictique "ce" avec l'objet sélectionné par la souris. La séquence d'actants obtenue après analyse sémantique permet d'activer et de compléter le script "CREER". Le dialogueur

doit ensuite être capable de rechercher dans son historique les informations utiles et manquantes, il doit être capable d'effectuer sur les informations disponibles les tests nécessaires à leur validation et gérer les diverses situations possibles (conflits, redondance, etc.).
Ce type de stratégie (ascendante) peut être combiné avec une stratégie descendante, lorsque les plans de l'utilisateur ont été inférés au préalable. Le script choisi doit s'inscrire dans un scénario actif et faire progresser ainsi la réalisation d'un plan. S'il est en contradiction avec les intentions de l'utilisateur, le système peut:
- soit remettre en cause les intentions de l'utilisateur,
- soit proposer au dialogueur une autre interprétation de la commande, en rectifiant au besoin les résultats de la reconnaissance de la parole,
- soit engager un dialogue avec l'utilisateur (Siroux, 1989).

5. CONCLUSION

ICPplan ouvre des perspectives pour la communication multimodale que nous croyons possible dès à présent malgré les performances encore limitées de la reconnaissance de la parole. Ce type de communication pose des problèmes nouveaux qui ne seront résolus qu'avec des concepts et des techniques nouvelles, en particulier pour la gestion des événements multimodaux.

Alors que les machines et les outils traditionnels étaient destinés à opérer sur des objets physiques, l'ordinateur est essentiellemnt destiné à manipuler l'information. Le champ de l'ergonomie traditionnelle avait pour but d'adapter la machine aux habiletés perceptivo-motrices de l'humain. L'ergonomie cognitive doit maintenant inclure les fonctions cognitives de l'usager pour une meilleure interaction avec l'ordinateur.
En effet, si la variété des domaines d'application de l'informatique n'a cessé de croître, le nombre des utilisateurs a suivi la même évolution. Or, peu d'interfaces intègrent, encore à l'heure actuelle, des concepts qui pourraient rendre ces interfaces plus ergonomiques donc plus efficaces et plus fiables. L'interaction multimodale participera dans un proche avenir à l'évolution des interfaces homme-machine

6. BIBLIOGRAPHIE

COUTAZ J., Interface homme-ordinateur : conception et réalisation. Dunod éd., Paris, 1990.

COUTAZ J. et CAELEN J., PRC communication homme-machine : Opération de Recherche Concertée interface homme-machine multimodale. Publication du PRC "Communication Homme-Machine", Juin 1990.

CAELEN J., Multimodal Interaction: Event Management and Experiments with ICPdraw. Pre-Proceedings of the second Venaco Workshop on Multimodal Dialogue, Maratea, Sept 1991.

COHEN Ph.R., On knowing what to say : Planning speech acts. Ph.D. Thesis, Technical Report n°118, Department of Computer Science, University of Toronto, January 1978.

COHEN Ph.R. et PERRAULT C.R., Elements of a Plan-Based Theory of Speech Acts. Cognitive Science 3, pp. 177-212, 1979.

FALZON P., Ergonomie Cognitive du Dialogue. PUG, Grenoble, 1990.

FALZON P., Multi-modal Interactions in MMI^2 Design Dialogues. Pre-Proceedings of the Second Venaco Workshop on Multimodal Dialogue, Maratea, Sept 1991.

FILLMORE C.J., The Case For Case. Bach E. and Harms R. eds, "Universals in Linguistic Theory", Holt, Rinehart and Wiston, pp 1-90, New York, 1968.

SIROUX J., GILLOUX M., GUYOMARD M., SORIN C., Le dialogue homme-machine en langue naturelle : un défi ? Annales des télécommunications, 44, n°1-2, 1989.

TAYLOR M.M., NEEL F., BOUHUIS D.G., The Structure of Multimodal Dialogue. Elsevier Science Publishers B.V., North-Holland, 1989.

VALOT C., AMALBERTI R., Description et analyse de l'activité de l'opérateur. Ecole IHM-M, Ecole Centrale, Lyon avril 1991.

Workshop IHM'91, groupe de travail interfaces multimodales, Dourdan, dec 1991.

19es J.E.P. - Bruxelles, 19 au 22 mai 1992.

Représentation structurelle du dialogue oral homme–machine et prédictions

Jean-Yves MAGADUR

CAP GEMINI INNOVATION
118, rue de Tocqueville, 75017 Paris, FRANCE.

Résumé

Ce papier décrit une représentation structurelle calculatoire du dialogue oral personne-machine orienté-tâche. Cette représentation est issue de l'adaptation du modèle de Roulet et Moeschler (Moeschler 89) utilisée dans le premier prototype du module de dialogue du projet Esprit SUNDIAL [1] (Bilange 91a), et est à la base de l'implémentation actuelle de ce module. Dans le premier prototype, on fait usage d'une structure de données, appelée structure de dialogue, qui constitue un historique hiérarchisé des actes de dialogue produits par les deux participants (l'utilisateur et le système) depuis le début du dialogue jusqu'à son état courant. Nous proposons d'inclure dans la structure de dialogue non seulement les actes déjà produits, mais aussi les prédictions du système sur les continuations possibles du dialogue. Nous montrons que l'intérêt de cette structure de dialogue enrichie réside dans la simplicité de sa mise à jour à chaque tour de parole.

1 Introduction

Le premier prototype du module de dialogue du projet SUNDIAL réalise l'implémentation du modèle présenté dans (Bilange 91a). Dans cette implémentation, on associe à un dialogue ou à un fragment de dialogue, une structure de données qui regroupe les actes de dialogue en interventions et en échanges. Cette structure de données sera appelée dans la suite structure de dialogue. Le modèle fournit des règles de syntaxe qui définissent ce qu'est une structure de dialogue valide. Comme elles sont utilisées pour la gestion dynamique du dialogue, ces règles de syntaxe doivent aussi rendre compte de la façon dont les structures représentant des états consécutifs du dialogue s'enchaînent dans le temps.

Le système de règles proposé dans le premier prototype présente l'inconvénient de ne pas se présenter sous la forme classique d'une grammaire de réécriture, utilisant des symboles terminaux et non-terminaux. De plus, ces règles conduisent à une procédure de mise à jour de la SD qui peut s'avérer complexe. En effet, pour tenir compte du caractère non déterministe du dialogue, on peut être amené, dans la mise à jour de la SD, à supprimer des informations précédement écrites pour les remplacer par de nouvelles. Autrement dit, la mise à jour de la SD ne s'effectue pas de façon incrémentale.

Notre propos est de montrer qu'en utilisant une structure de dialogue plus riche, qui contient non seulement un historique des actes de dialogue déjà produits mais aussi des indications sur les continuations possibles (c'est à dire des prédictions), on peut s'affranchir des ces deux inconvénients. Cette structure de dialogue enrichie sera appelée *structure de dialogue dynamique*. L'algorithme que nous proposons est implémenté dans la version actuelle du module de dialogue du projet SUNDIAL.

2 Présentation du modèle

2.1 Les actes de dialogue

Nous définissons les actes de dialogue à partir de la notion plus primitive d'état mental (Searle 83) (Sadek 91). A tout instant, l'état mental du système est un ensemble de prédicats qui représentent les croyances courantes du système (pour nous, les intentions du système sont aussi des croyances de celui-ci). Un acte de dialogue se

[1] Ce projet est partiellement financé par le programme ESPRIT de la communauté européenne. Les partenaires du projet sont CAP GEMINI INNOVATION, CNET, CSELT, DAIMLER-BENZ, UNIVERSITÉ D'ERLANGEN, INFOVOX, IRISA, LOGICA, POLITECHNICO DI TORINO, SARITEL, SIEMENS, UNIVERSITÉ DE SURREY

déclenche lorsque certaines croyances qui sont les préconditions de l'acte, font partie de l'état mental courant. La production d'un acte a pour effet de modifier l'état mental du système, c'est-à-dire de supprimer certaines croyances et d'en ajouter de nouvelles. L'état mental du système évolue sous l'effet de deux facteurs extérieures : les énoncés de l'utilisateur (traduits dans un langage de représentation sémantique) et les messages du module de tâche. (La séparation entre module de dialogue et module de tâche, réalisée dans un but de généricité, est décrite dans (Bilange 91b)). Un message du module de tâche peut être, par exemple, une requête demandant la valeur d'un paramètre de la tâche tel que, dans l'application française du projet, l'heure de départ ou le jour de départ d'un avion.

Nous utiliserons dans la suite la notion d'appariement des actes de dialogue telle qu'elle est définie dans (Bilange 91a). L'idée intuitive est qu'une question est appariée avec ses réponses possibles (ou tout au moins avec les plus probables, c'est à dire avec celles qui sont observées le plus souvent dans un corpus de dialogues de référence), qu'une réponse est appariée avec les différents types de confirmation qu'on peut lui associer, etc. On peut donc, en se fondant sur une liste d'actes appariés, faire à tout instant des prédictions sur les continuations possibles d'un dialogue.

2.2 Les échanges

Dans le modèle dont nous sommes partis, les échanges sont constitués d'interventions, elles-même constituées d'actes de dialogue. Pour simplifier l'exposé nous supposerons dans la suite que les interventions des locuteurs se composent toujours d'exactement un acte. Nous confondrons donc interventions et actes de dialogue. Ainsi, pour nous, une SD est une structure de données qui contient les labels des actes produits au cours du dialogue, regroupés par échanges. Les actes de dialogue ont déjà été discutés au paragraphe précédent. Les échanges servent à réaliser soit un transfert d'information, soit une clarification. Le prototype de l'échange est un couple question–réponse, ou bien un couple question–réponse suivi d'une confirmation. Considérons par exemple le fragment de dialogue suivant :

S_1 *Quel jour voulez-vous partir ?*
U_1 *Le 6 juin.*
S_2 S_2^1 *Le 6 juin.*
S_2^2 *A quelle heure ?*
U_2 *A huit heures.*

La séquence (S_1, U_1, S_2^1) definit un premier échange E_1, et (S_2^2, U_2) en définit un deuxième. La question S_1, qui est le premier acte de l'échange E_1 est *l'initiative* de l'échange. La réponse U_1 est la *réaction.* Enfin, on dit que la confirmation S_2^1 est l'évaluation de l'échange. Un locuteur produit une évaluation pour exprimer sa satisfaction ou sa désaprobation à propos de l'échange. On dit qu' initiative, réaction et évaluation sont les trois fonctions illocutoires possibles d'un acte de dialogue. La SD associée à ce fragment de dialogue s'écrit donc :

$$D \left[\begin{array}{ll} E & \left[\begin{array}{ll} S_1 & \textit{Quel jour voulez – vous partir?} \\ U_1 & \textit{Le 6 juin.} \\ S_2^1 & \textit{Le 6 juin.} \end{array} \right. \\ E & \left[\begin{array}{ll} S_2^2 & \textit{A quelle heure?} \\ U_2 & \textit{A huit heures.} \end{array} \right. \end{array} \right.$$

Figure 1: Deux échanges consécutifs

Une définition opérationelle de la notion d'échange doit fournir des critères permettant de décider, lorsqu'un acte est produit par un locuteur, si on doit placer cet acte à l'intérieur d'un échange déjà existant (et si oui, à quelle place), ou bien ouvrir un nouvel échange. La grammaire de dialogue que nous définissons à la section suivante constitue une telle définition opérationelle. Nous nous en tenons pour l'instant, à la conception intuitive qui vient d'être présentée.

Afin de préciser la notion d'échange, nous allons examiner quelques exemples qui nous conduiront à la construction d'une grammaire définissant les SD valides et qui sera à la base de la définition des structures de dialogue dynamiques. Considérons d'abord le dialogue suivant :

S_1 *Quel jour voulez-vous partir ?*
U_1 *Le 6 juin.*
S_2 *Le 6 juin ?*
U_2 *Oui, c'est ça, le 6 juin.*

U_2 est visiblement la réaction à S_2. S_2 et U_2 constituent donc un échange, et ceci nous conduit naturellement à l'idée que l'évaluation d'un échange peut être elle–même un échange, qui sera alors dit échange évaluateur. Ceci nous mène à la SD présentée figure 2.

$$D \left[\begin{array}{ll} E & \left[\begin{array}{ll} S_1 & \textit{Quel jour voulez – vous partir?} \\ U_1 & \textit{Le 6 juin.} \\ E & \left[\begin{array}{ll} S_2 & \textit{Le 6 juin?} \\ U_2 & \textit{Oui, c'est ça, le 6 juin} \end{array} \right. \end{array} \right. \end{array} \right.$$

Figure 2: Un exemple d'échange évaluateur

Nous verrons sur des exemple ultérieurs qu'il peut y avoir des échanges évaluateurs imbriqués. On en

arrive ainsi à la conception que l'évaluation d'un échange est parfois un acte de dialogue (figure 1) et parfois un échange (figure 2). On peut cependant adopter une vision plus unitaire en convenant que l'évaluation d'un échange est dans tous les cas un échange. Dans certains cas, cet échange évaluateur ne sera constitué que d'un seul acte alors que dans d'autres il en comprendra au moins deux. Avec cette convention, la SD de la figure 1 est, en toute rigueur, incorrecte. En effet, l'évaluation S_2^1 devrait être incluse dans un échange. On peut néanmoins continuer à utiliser cette représentation, en la considérant comme une notation abrégée de la SD correcte. Examinons maintenant le dialogue suivant :

S_1 *En quelle classe ?*
U_1 *Quel est le prix en première classe ?*
S_2 *2700 francs.*
U_2 *En première, alors.*

U_2 est clairement la réaction à l'initiative S_1. On constate qu'un échange vient s'insérer entre l'initiative et la réaction. Cet échange, que nous appellerons échange incident, a un statut particulier. En l'ouvrant, l'utilisateur commence l'exécution d'un plan dont le but est de répondre à la question du système (à laquelle il n'est pas capable de répondre immédiatement, par manque d'information). L'échange incident constitue donc une phase préliminaire à la réaction proprement dite. Entre l'initiative et la réaction, il peut y avoir un nombre quelconque d'échanges incidents. La figure 3 représente la SD associée à ce dialogue.

$$D\left[\begin{array}{ll} E & \left[\begin{array}{ll} S_1 & \textit{En quelle classe?} \\ E & \left[\begin{array}{ll} U_1 & \textit{Quel est le prix} \\ & \textit{en premiere classe?} \\ S_2 & 2700\ \textit{francs.} \end{array}\right. \\ U_2 & \textit{En premiere, alors.} \end{array}\right. \end{array}\right.$$

Figure 3: Un exemple d'échange incident

Les exemples que nous avons étudiés jusqu'à présent nous conduisent à conclure qu'un échange est constitué d'un acte de dialogue, suivi d'une suite éventuellement vide d'échanges incidents, puis d'une réaction optionnelle et d'un échange évaluateur optionnel. Cette description se simplifie considérablement si on convient de dire que la réaction est elle–même un échange (réduit à un seul acte). On peut alors énoncer le principe suivant, qui régit la forme des structures de dialogue, c'est-à-dire la façon dont les échanges s'emboîtent dans une SD valide :

Un échange est constitué d'un acte de dialogue suivi d'une suite éventuellement vide d'échanges

L'adoption de cette nouvelle convention rend à nouveau, en toute rigueur, les figures précédentes incorrectes, puisque les réactions devraient y figurer en étant incluses dans un échange. On peut cependant là encore, comme pour la convention précédente (à propos des évaluations), les considérer comme des notations abrégées des SD correctes.

Des conventions de représentation que nous venons d'adopter il résulte que toute production d'un acte de dialogue par un des deux locuteurs se traduit par l'ouverture d'un échange (puisque les évaluations et les réactions sont systématiquement des échanges, éventuellement réduits à un seul acte). Dans ces conditions, il est naturel d'associer à un échange la fonction illocutoire de l'acte de dialogue par lequel il est ouvert. Il y aura donc des échanges à fonction d'initiative, de réaction, ou d'évaluation, et des échanges incidents, dont le rôle est de préparer une réaction et auxquels nous n'attribuerons pas de fonction illocutoire. Un échange est à fonction d'initiative s'il n'est pas sous–échange d'un autre échange. L'acte par lequel il est ouvert introduit un changement de contexte dans le dialogue. Les actes qui peuvent ouvrir un échange à fonction d'initiative forment une liste que nous noterons L_i dans la suite.

2.3 La structure de dialogue

Les considérations précédentes conduisent naturellement à une grammaire qui rend compte de la forme des SD valides. On introduit, à cette fin, un non–terminal $\mathcal{E}$ dont la fonction est d'être réécrit en une suite d'échanges. La SD dynamique sera une production de cette grammaire telle que, si on supprime tous les non–terminaux qu'elle contient, on obtienne la SD qui représente l'état courant du dialogue. Comme un dialogue est constitué d'une suite d'échanges, la SD dynamique avant le début du dialogue s'écrit :

$$D[\mathcal{E}$$

Le principe régissant l'emboîtement des échanges dans les SD valides se traduit par les règles de réécriture :

- $\mathcal{E} \rightarrow \begin{array}{l} E \\ \mathcal{E} \end{array} \left[\begin{array}{l} \textit{acte} \\ \mathcal{E} \end{array}\right.$
- $\mathcal{E} \rightarrow \varepsilon$

où *acte* décrit l'ensemble des labels d'actes de dialogue et où ε est la SD vide. La première règle indique qu'une suite d'échanges non vide est composée d'un échange suivi d'une liste d'échanges et

que tout échange est constitué d'un acte suivi d'une liste d'échanges. La deuxième règle indique qu'une suite d'échanges peut être vide.

Cette grammaire ne permet pas de distinguer entre les différentes fonctions illocutoires possibles d'un échange. De plus, elle ne tient aucun compte de l'appariement des actes de dialogue, ni même du fait que l'initiative et la réaction d'un échange doivent être produits par des locuteurs différents.

Si on veut obtenir une grammaire qui rende compte non seulement de l'emboîtement des échanges, mais aussi de leurs fonctions illocutoires, il faudra paramétrer le non-terminal $\mathcal{E}$ par une fonction illocutoire. La prise en compte des actes de dialogue proprement dits nécessite, quant à elle, de paramétrer les non-terminaux par des listes d'actes. Ainsi, si l'acte de dialogue a vient d'être produit, et qu'il est apparié avec chacun des actes $a_1, \ldots, a_n$, la SD dynamique va contenir un non-terminal paramétré par la liste $(a_1, \ldots, a_n)$. Ceci signifiera que si dans un tour de parole ultérieur l'un des actes de la liste, disons a_i, est déclenché, le non-terminal sera réécrit, et inscrira dans la SD dynamique l'acte a_i. Enfin, dans un même échange, les contextes des différents actes sont sémantiquement voisins (ce qui permet d'associer à un échange ce que Grosz et Sidner appèlent un espace attentionel (Grosz & Sidner 86). Pour tenir compte de ce fait, les non-terminaux générateurs de réactions et d'évaluations devront comporter une contrainte sur le contexte. Ces idées nous conduisent à la grammaire suivante :

- $\mathcal{E}(\mathrm{i}, L_i) \rightarrow E\left[\begin{array}{l} acte \\ \mathcal{E}(\mathrm{r}, \mathrm{L'}) \end{array}\right.$
  $\mathcal{E}(\mathrm{i}, L_i)$
- $\mathcal{E}(\mathrm{r}, \mathrm{L}) \rightarrow E\left[\begin{array}{l} acte \\ \mathcal{E}(\varepsilon) \end{array}\right.$
  $\mathcal{E}(\mathrm{e}, \mathrm{L'})$
- $\mathcal{E}(\mathrm{e}, \mathrm{L}) \rightarrow E\left[\begin{array}{l} acte \\ \mathcal{E}(\mathrm{r}, \mathrm{L'}) \end{array}\right.$
  $\mathcal{E}(\varepsilon)$
- $\mathcal{E}(\varepsilon) \rightarrow \varepsilon$

Pour simplifier les notations, nous ne faisons pas figurer les contraintes de contexte comme paramètres des non-terminaux. Dans toutes les règles de réécriture, *acte* appartient à la liste paramétrant la partie gauche de la règle, L' est la liste des actes de dialogue appariés avec *acte*, et i, r et e désignent les trois fonctions illocutoires. Cette grammaire rend compte des phénomènes de dialogue illustrés par les exemples précédents, à l'exception de celui des échanges incidents. Pour rendre compte de ce phénomène, il faut être capable de reconnaître si un acte ouvre un échange incident ou s'il ouvre un échange à fonction d'initiative. On peut résoudre ce problème en remarquant que le contexte d'un échange incident et le contexte de l'échange auquel il appartient doivent être sémantiquement voisins. La difficulté est bien sur de disposer d'une définition suffisamment précise de cette proximité sémantique. Une autre solution consiste à munir le système de capacités d'inférence de plans et de détection d'obstacles dans un plan.

Voyons de quelle façon on utilise cette grammaire pour la gestion dynamique du dialogue. Lorsque l'état mental du système est modifié par un facteur extérieur (par exemple un message du module de tâche ou une entrée de l'utilisateur), le système entame un parcours récursif de la SD dynamique courante. Lorsqu'il rencontre un non-terminal, il vérifie si l'un des actes associés à ce non-terminal est déclenchable. Si c'est le cas, cet acte est déclenché, et le non-terminal est réécrit; sinon, le parcours récursif continue. Cette procédure est la même pour l'interprétation des actes de l'utilisateur et pour la génération des actes du système.

2.4 Un exemple

Afin d'illustrer cet algorithme, nous allons montrer l'évolution de la structure dynamique au début du dialogue de la figure 1. Avant le début du dialogue, la SD dynamique s'écrit :

$$D[\mathcal{E}(\mathrm{i}, L_i)$$

A cet instant, le module de dialogue reçoit une requête du module de tâche, qui demande le jour de départ. Ceci modifie l'état mental du système, et permet de déclencher l'acte de dialogue *question_ouverte*. Cet acte est apparié avec l'acte *informe* de l'utilisateur. La SD dynamique devient donc :

$$D\left[\begin{array}{l} E \\ \mathcal{E}(\mathrm{i}, L_i) \end{array}\right.\left[\begin{array}{l} \textit{Quel jour voulez} - \textit{vous partir?} \\ \mathcal{E}(\mathrm{r}, (\ldots, \textit{informe}, \ldots)) \end{array}\right.$$

L'énoncé suivant de l'utilisateur modifie l'état mental de telle sorte qu'il est possible de déclencher l'acte *informe*. Il est par contre impossible de déclencher les actes de la liste L_i. L'acte *informe* est apparié avec l'acte *confirme*, si bien que la nouvelle SD dynamique s'écrit :

$$D\left[\begin{array}{ll} E & \left[\begin{array}{l} \textit{Quel jour voulez – vous partir?} \\ E\left[\begin{array}{l}\textit{Le 6 juin}\\ \mathcal{E}(\varepsilon)\end{array}\right] \\ \mathcal{E}(\text{e, (..., } \textit{confirme}\text{, ...))}\end{array}\right. \\ \mathcal{E}(\text{i, } L_i) & \end{array}\right.$$

La réponse de l'utilisateur modifie l'état mental de telle sorte que l'acte *confirme* soit déclenchable. Lorsque la réponse est reçue, le module de dialogue l'envoie au module de tâche afin que celui–ci vérifie la cohérence de l'information et émette éventuellement de nouvelles requêtes. Le module de tâche envoie alors une requête sur l'heure de départ, ce qui rend l'acte *question_ouverte* déclenchable. Deux actes sont donc déclenchables. Au cours du parcours récursif de la SD dynamique, on rencontre d'abord le $\mathcal{E}(\varepsilon)$, qui est remplacé par ε, puis le $\mathcal{E}$(e, (..., confirme, ...)), qui est déclenché, et enfin le non–terminal $\mathcal{E}$(i, L_i) qui lui aussi est déclenché. On aboutit donc à la SD dynamique suivante :

$$D\left[\begin{array}{ll} E & \left[\begin{array}{l} \textit{Quel jour voulez – vous partir?} \\ E\left[\begin{array}{l}\textit{Le 6 juin}\\ E\left[\begin{array}{l}\textit{Le 6 juin}\\ \mathcal{E}(\text{r, (...)})\end{array}\right]\end{array}\right. \\ \mathcal{E}(\varepsilon)\end{array}\right. \\ E & \left[\begin{array}{l}\textit{A quelle heure?}\\ \mathcal{E}(\text{r, (..., } \textit{informe}\text{, ...))}\end{array}\right. \\ \mathcal{E}(\text{i, } L_i) & \end{array}\right.$$

A l'étape suivante, la réponse de l'utilisateur déclenchera le non–terminal $\mathcal{E}$(r, (..., informe, ...)).

Cet exemple illustre la simplicité de l'algorithme de mise à jour de la SD dynamique, qui se réduit à un simple parcours récursif. Un autre avantage de cet algorithme est que le calcul des prédictions se fait en même temps que cette mise à jour de la SD.

Le dialogue suivant constitue un autre exemple important :

S_1 *Quel jour voulez-vous partir ?*
U_1 *Le 6 juin.*
S_2 S_2^1 *Le 6 juin.*
S_2^2 *A quelle heure ?*
U_2 *Non, le 10 juin.*
S_3 *Le 10 juin ?*
U_3 *Oui.*
S_3 *A quelle heure ?*
U_3 *A huit heures.*

Avec U_2, l'utilisateur poursuit l'échange évaluateur ouvert par S_2^1 et ne répond pas à l'initiative S_2^2. Le système devra donc réintroduire cette initiative. Mais pour cela, il doit attendre la fermeture de l'échange évaluateur. L'état mental du système doit pour cela contenir un prédicat indiquant l'absence de réponse de l'utilisateur et un autre prédicat indiquant qu'une tentative de clarification est en cours. En ne répondant pas à S_2^2, l'utilisateur déclenche un acte de dialogue que nous avons noté *aucune_réponse(question_ouverte)* sur la figure. Cet acte est la réaction à S_2^2. L'intérêt d'une telle représentation est qu'il semble alors homogène de considérer la réintroduction comme l'évaluation de l'échange, comme ceci est indiqué sur la figure 4.

$$D\left[\begin{array}{ll} E & \left[\begin{array}{l}\textit{Quel jour voulez vous partir?}\\ \textit{Le 6 juin}\\ E\left[\begin{array}{l}\textit{Le 6 juin?}\\ E\left[\begin{array}{l}\textit{Non, le 10 juin.}\\ E\left[\begin{array}{l}\textit{Le 10 juin?}\\ \textit{Oui.}\end{array}\right]\end{array}\right]\end{array}\right.\end{array}\right. \\ E & \left[\begin{array}{l}\textit{A quelle heure?}\\ \textit{aucune\_réponse(question\_ouverte)}\\ E\left[\begin{array}{l}\textit{A quelle heure?}\\ \textit{A huit heures.}\end{array}\right]\end{array}\right. \end{array}\right.$$

Figure 4: Réintroduction d'une initiative

3 Un algorithme équivalent

Le rôle du module de dialogue est de décider, sur la base des informations qu'il reçoit de l'extérieur (les messages du module de tâche et les énoncés de l'utilisateur), quels doivent être les actes de dialogue produits par le système à chaque tour de parole. Dans l'algorithme que nous avons décrit, on peut noter que seuls les non–terminaux de la SD dynamique sont utilisés pour cette prise de décision. En effet, c'est uniquement sur les symboles non–terminaux qu'on s'arrête au cours du parcours récursif de la SD dynamique. Ceci signifie que l'algorithme que nous avons présenté peut être modifié en supprimant tout recours aux terminaux de la SD dynamique.

Il faut pour cela modifier la grammaire de dialogue de telle sorte que les membres de droite des règles de réécriture ne contiennent plus que des non–terminaux. Ceci conduit aux règles suivantes :

- $\mathcal{E}$(i, L_i) → $\mathcal{E}$(r, L'), $\mathcal{E}$(i, L_i)
- $\mathcal{E}$(r, L) → $\mathcal{E}$(e, L')
- $\mathcal{E}$(e, L) → $\mathcal{E}$(r, L')

Quant à la valeur initiale de la SD dynamique (avant le début du dialogue), elle devient :

($\mathcal{E}$(i, L_i))

Alors, à tout instant, la SD dynamique sera une liste de non–terminaux. Comme les non–terminaux sont paramétrés par les prédictions courantes, on peut voir cette SD dynamique comme une liste ordonnée des prédictions courantes. Dans le nouvel algorithme fondé sur l'utilisation de cette grammaire, on n'a plus recours à l'historique des actes produits au cours du dialogue. Seules sont utilisées les prédictions courantes du système, qui sont intégrées à l'état mental. Ce sont ces prédictions qui sont essentielles pour le fonctionnement de l'algorithme. Cependant, il faut noter que le moyen par lequel elles sont calculées importe peu. Nous avons supposé, pour la commodité de l'exposé, que le calcul des prédictions se faisait à l'aide de paires d'actes de dialogue, mais il est clair que toute autre méthode de calcul peut convenir.

4 Conclusion

Notre algorithme, en substituant aux structures de dialogue décrites dans (Bilange 91a) des structures plus complexes (les SD dynamiques), permet en retour de disposer d'une méthode plus simple pour la mise à jour de ces structures. L'opération de mise à jour de la SD dynamique se réduit en effet à un simple parcours récursif de cette structure. Le modèle a été initialement présenté en faisant l'hypothèse que le calcul des prédictions se fondait sur l'utilisation de paires d'actes de dialogue. Cependant, comme nous le remarquons dans la dernière section, la grammaire de dialogue est en fait indépendante de la façon particulière dont sont calculées les prédictions, pourvu que ces prédictions existent. D'autres méthodes de calcul des prédictions fondées sur des modèles plus fin de l'utilisateur sont donc a priori compatibles avec notre algorithme.

Remerciements

Je voudrais remercier particulièrement David Sadek qui, outre l'aide qu'il m'a apportée pour la relecture de cet article, a aussi motivé, par nos conversations, certaines idées présentées dans ce papier, en particulier la représentation proposée pour la réintroduction d'initiatives.

Références

Bilange, E. (1991a), Modélisation du dialogue oral finalisé personne-machine par une approche structurelle, théorie et réalisation. Thèse de l'Université de Rennes 1, Décembre 1991.

Bilange, E. (1991b), A task independent oral dialogue model. In proceedings of the *European chapter of the ACL*, April, Berlin.

Grosz, B.J. & Sidner, C.L. (1986), Attention, Intentions, and the structure of discourse. *Computational Linguistics*, Vol. 12, No 3, July–September.

Moeschler, J. (1989), *Modélisation du dialogue, représentation de l'inférence argumentative.* Hermès.

Sadek, M.D. (1991), *Attitudes mentales et interaction rationnelle : vers une théorie formelle de la communication*, Thèse de l'Université de Rennes 1, Juin 1991.

Searle, J.R. (1983), *L'Intentionnalité - Essai de philosophie des états mentaux*, Les Editions de Minuit, Paris, 1983.

UN SUPERVISEUR INTELLIGENT POUR LA GESTION DES CONNAISSANCES LINGUISTIQUES EN RECONNAISSANCE DE LA PAROLE.

THIERRY SPRIET

LABORATOIRE D'INFORMATIQUE
DE L'UNIVERSITE D'AVIGNON (FRANCE)

Résumé

SYRAPAC is a continuous speech recognition system, with explicit knowledge use orientation. Its great independence between data and treatments permits an easy adaptation to new applications. SYRAPAC is a multi-expert system managed by a supervisor which controls both the strategy and the process of each expert. The general algorithm, A* and "best first" stems from "the Island Driven Strategy" (WOODS, 1982). It makes partial, bottom-up linguistic parsing. Whenever the supervisor studies the situation and chooses an appropriate algorithm according to the capacity of each expert. Knowledge is uniformaly formalised in such a way that the supervisor could use it independently of his origin (lexicon, syntax,...)

1. Introduction

Nous présentons dans cet article la structure générale et les mécanismes de base de SYRAPAC, SYstème de Reconnaissance Automatique de la PArole Continue. Notre système est basé sur l'utilisation de connaissances explicites, extraites de divers experts (acoustico-phonétique, lexical, syntaxique). La gestion de ces connaissances est réalisée par un superviseur utilisant des techniques formelles de résolution de systèmes de contraintes. Ceci nous donne une grande indépendance vis-à-vis des connaissances, autorisant ainsi aisément leur évolution, modification ou substitution, mais aussi l'intégration de connaissances d'autres experts (prosodique, sémantique, phonologique ou pragmatique). SYRAPAC utilise un algorithme d'analyse par îlots de confiances, dérivé de "Island Driven Stratégie" (WOODS,1982). Le superviseur est chargé de la distribution des tâches aux experts et de l'analyse globale de la situation. Son raisonnement sur un plan général lui permet de déterminer à chaque étape les actions les plus opportunes en fonction des capacités de chaque expert.

2. Les experts

Le système utilise deux types d'experts, des expert autonomes, et des experts contrôlés.

2.1. LES EXPERTS AUTONOMES

Leurs procédures d'analyse ne sont pas sous contrôle du superviseur qui ne fait que choisir l'instant de leur activation et se contente d'exploiter leurs résultats. Ils sont au nombre de deux :

- Le module de Décodage Acoustico-Phonétique, qui a deux fonctionnalités, initialisation du treillis phonétique et vérification d'hypothèses émises par les niveaux supérieurs. Lors de sa phase ascendante (initialisation) il détecte sur le signal des hypothèses phonétiques auxquelles il attribue un taux de vraisemblance. Ces éléments sont les îlots de base de l'algorithme. En phase descendante, le superviseur le contacte pour vérification et évaluation de la vraissemblance de prédictions phonétiques émises par les autres experts.
- Le module de dialogue inhérent à l'application elle-même. Il intervient en fin de processus

lorsque la solution la plus vraissemblable a été extraite de l'entrée vocale. Son rôle est d'interpréter la formule logique associée à la solution et d'effectuer le traitement désiré (recherche d'informations dans une base de données, commande d'un automate, etc). Si une réponse orale ou graphique est attendue, c'est ce module qui doit en prendre la charge.

2.2. LES EXPERTS SOUS CONTROLE

Ils sont décomposés en trois parties, un module de connaissance, un module de traitements formels, et un "moteur d'analyse". Le premier est une base de faits, constituant la connaissance d'un expert du domaine. Les modules de traitements formels effectuent des unifications sur des structures de données à parties variables. Le "moteur d'analyse" d'un expert est l'ensemble des règles de gestion de la connaissance, l'analyseur proprement dit. Ces règles sont intégrées au superviseur au sein même des algorithmes qu'il utilise pour les analyses linguistiques partielles effectuées lors de la jonction de deux îlots. Le superviseur possède en outre une interface formalisant les connaissances, sous forme de listes ou d'arbres d'identificateurs et de variables, lui permettant ainsi d'utiliser directement les fonctionnalités offertes par les modules de traitements formels.

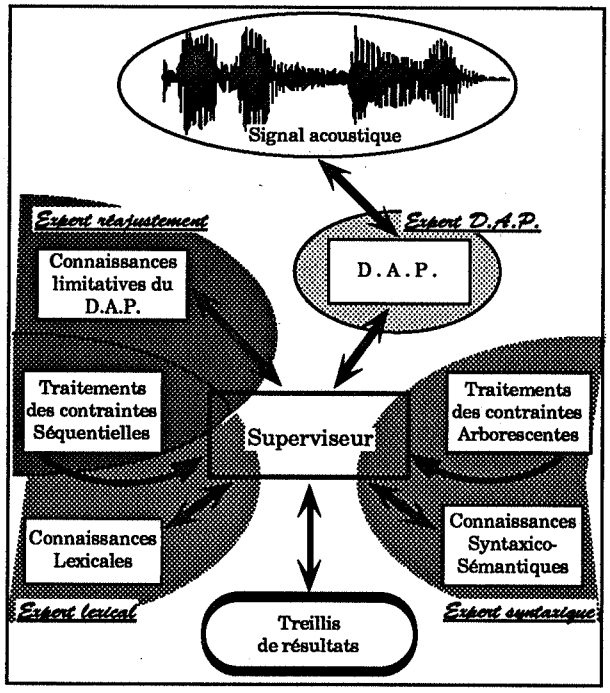


figure 1 Schéma général.

Le système intégre trois de ces experts.

• L'expert lexical, constitué de :
- un dictionnaire pouvant fournir la décomposition phonétique de toutes les formes lexicales, et leurs attributs syntaxico-sémantiques,
- un module de résolution de contraintes séquentielles, effectuant des unifications partielles sur des listes à parties variables.
- et un "moteur d'analyse" intégré dans le superviseur.

• L'expert syntaxique est constitué quant à lui de :
- une base de connaissances syntaxiço-sémantiques, donnant les "règles d'assemblage" des unités lexicales. Un prétraitement permet au système d'intégrer toute grammaire de métamorphose (COLMERAUER, 1982).
- un module de résolution de contraintes arborescentes,
- le "moteur d'analyse" intégré au superviseur.

• L'expert de réajustement des performances du module de D.A.P. est lui constitué de :
- un ensemble de règles de connaissances interprétant les "échecs" probables du D.A.P. Ces règles permettent d'effectuer des prédictions phonétiques pour des unités que le DAP n'a pas détectées, à cause d'un contexte phonologique difficile ou absorbant par exemple.
- le même module de résolution de contraintes séquentielles que l'expert lexical,
- le "moteur d'analyse" intégré au superviseur.

3. Algorithme général

Les modules de DAP actuellement opérationnels ne sont pas en mesure de fournir intégralement la chaîne phonétique effectivement prononcée, et on peut raisonnablement penser que ce but est utopique. Par contre ces modules peuvent estimer de façon satisfaisante la vraisemblance d'hypothèses phonétiques détectées dans le signal. Sur la base de ces scores, une stratégie basée sur la croissance d'îlots de confiance semble naturelle. Nous avons donc repris l'algorithme "Island-Driven Strategy" en isolant la base de données du module lexical, intégrant complètement son "moteur d'analyse" au super-

viseur, et en adaptant le calcul du score de priorité à notre treillis phonétique de départ qui n'est pas saturé, ni segmenté linéairement. Cette structure de treillis mémorise aussi les îlots en attendant leur traitement ; un îlot peut être un phonème (hypothèse phonétique détectée par le DAP) ou un groupe de phonèmes résultat d'une collision effectuée lors d'une étape antérieure.
La "confiance" accordée à chaque îlot est quantifiée par son score de priorité. Celui-ci est une densité du taux de vraisemblance de sa décomposition phonologique de l'îlot par rapport au signal. Ce taux est calculé sur les valuations des hypothèses phonétiques détectées par le DAP, et sur les résultats de l'analyse acoustico-phonétique descendante effectuée sur les prédictions des niveaux supérieurs.

L'algorithme est le suivant :

1 Recherche de l'îlot de plus fort score de priorité dans le treillis.
2 S'il recouvre la phrase et qu'il a une interprétation complète pour les modules syntaxique et lexical, c'est une solution. Une recherche identique est alors effectuée pour tous les îlots de même score. Toutes les solutions obtenues, sont déclarées résultats équivalents, et interprétées par le module de dialogue.
3 Sinon tous les îlots voisins sont énumérés, et le superviseur effectue (avec l'aide des experts) des tentatives de collision pour chacun. A chaque succès, un nouvel îlot est créé et inséré dans le treillis résultat.
4 L'îlot traité est retiré du treillis.
5 Retour en 1.

Le choix d'un tel algorithme, présuppose des capacités techniques capables de surmonter l'explosion combinatoire du nombre d'îlots engendrés. Si pour le moment, nous sommes pénalisés par un temps d'accès aux données trop lent, les développements actuels du matériel informatique nous laisse espérer un temps de réponse plus propice à un dialogue homme/machine dans un avenir proche.

4. Le superviseur

4.1. SCHEMA GENERAL

Outre le bon déroulement de l'algorithme de base, il est chargé des analyses linguistiques partielles lors des tentatives de collisions d'îlots. Les îlots sont classés en fonction de leur passé, à savoir des analyses dont ils ont déjà fait l'objet. Ces analyses sont faites lors de la jonction de deux îlots, afin de déterminer si leur union n'est pas en contradiction avec les connaissances linguistiques des divers experts.

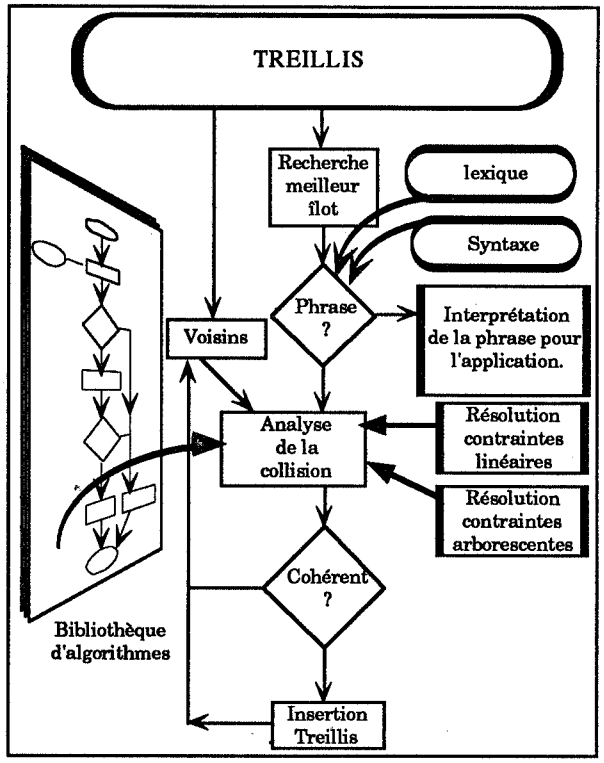


figure 2 Schéma du superviseur
Algorithme et flux de données

A chaque îlot nous associons des informations telles que : les phonèmes qu'il recouvre déjà et ceux qu'il attend, les unités lexicales et les catégories syntaxiques qui lui donnent une légitimité linguistique. Les informations déjà réalisées par l'îlot seront appelées couvertures (lexicale ou syntaxique) tandis que celles restant à réaliser seront des contraintes. Nous avons établi sept catégories permettant de classer les îlots en fonction des propriétés :

- avec ou sans contraintes phonétiques,
- avec ou sans contraintes syntaxiques,
- avec ou sans couverture lexicale,
- avec ou sans couverture syntaxique.

Les combinaisons incompatibles étant supprimées (sans couverture lexicale, avec contraintes phonétiques par exemple). Nous obtenons 7

classes distinctes. Lors d'une tentative de jonction d'îlots, 49 cas contextuels peuvent se produire. Pour chacun d'eux le superviseur accède à un algorithme spécifique de résolution. Ces algorithmes font appels aux traitements formels de résolution de contraintes, et travaillent sur des données symboliques. Il sont regroupés dans une bibliothèque partie intégrante du superviseur.

4.2.LES SYSTEMES DE CONTRAINTES

Un îlot est composé de divers types d'informations :

• des informations sur sa composition "physique", telles que la zone temporelle qu'il recouvre, la liste des sons qui le composent, un score de vraisemblance, une zone de stabilité.

• Des informations sur sa composition et ses prédictions syntaxiques, lexicales et phonétiques. Les informations syntaxiques sont sauvegardées sous forme d'arbres à parties variables, Tandis que les informations lexicales et phonétiques sont des listes de compositions et de prédictions. Ces structures sont organisées pour mettre en évidence la présence ou l'absence de prédictions, ceci permettant au superviseur de classifier chaque îlot en scrutant uniquement ces structures.

Exemple d'un îlot et des informations associées

```
ILOT(1548, <65,158>, <60,151>) ->;
/* donne le numéro d'identification de l'îlot, sa zone
de couverture temporelle et sa zone de stabilité (zone
de plus grande vraisemblance)
*/
SCORE_ILOT(1548, 0.56 , 80) ->;
/* donne le taux de vraisemblance de l'îlot, et son
score de priorité. Ils sont recalculés à chaque nouvelle
extention.
*/
COMPOSITION_PHON(1548,
        kk.ou.ai.ii.pp.nil) ->;
/* contient la décomposition phonétique recouverte
par l'îlot.
*/

STRUCT_LEX(1548,
        <decouvert( verb_pp.
                x_genre. x_nombre.nil),
          dd.ei.nil,
          kk.ou.(vv).ai.(rr).nil,
          nil>.
        <hyperion(nom_p.sin.mas. planete.nil),
          nil,
          ii.pp.nil,
          ai.rr.yy.on.nil>.
        nil) ->;

/* Liste des mots que recouvre l'îlot. les trois listes de
phonèmes associées à chaque entité lexicale, déterminent les phonèmes déjà reconnus (2° liste) et ceux
restant à recouvrir à droite et à gauche, ceux sont les
contraintes lexicales (1° et 3° liste).
*/
STRUCTURE_SYN(1548,
        <aux2(p_as.x_genre. x_nombre.nil).nil,
         verb2(pp. x_genre. x_nombre .nil).
         <compl_verb2( x_genre. x_nombre.nil)
           , <npropre (x_genre.x_nombre.nil)>.
          nil>,
        nil>.nil) ->
contr(eq(x_genre,mas).eq(x_nombre,sin).
dif(p_as,etre).nil);
/* A chaque îlot on associe un arbre, où les branches
droite et gauche sont des prédictions, elles ne sont pas
développées ; tandis que la branche centrale est la
liste des catégories déjà recouvertes ou partiellement
recouvertes. Chaque noeud de cette dernière est un
arbre du même type que celui décrit ici.
*/
```

Les trois premières règles sont les informations concernant la composition physique de l'îlot, elles n'interviennent dans l'analyse que pour choisir l'îlot de meilleur score, et pour déterminer les unitésvoisines temporellement.

Les deux règles suivantes (structure lexicale, structure syntaxique) sont les résultats des analyses linguistiques déjà effectuées sur l'îlot. Ce sont elles qui forment le système de contraintes associé à l'îlot.

Lors de la collision de deux îlots le superviseur tente de faire fusionner les deux systèmes. La jonction est acceptée si le système résultant est cohérent, et n'est pas formé de deux sous-systèmes indépendants.

4.3. RESOLUTION D'UN SYSTEME

Chaque système propose un ensemble d'éléments déjà recouverts par l'îlot (sons, unités lexicales, ou catégories syntaxiques), ainsi que des prédictions, unités restant à recouvrir pour valider complètement les structures lexicale et syntaxique. Lors d'une jonction d'îlots, les prédictions de l'un sont des contraintes que le superviseur va essayer d'unifier avec la décomposition du second. Il utilise pour cela les modules de traitements formels à qui il fournit des structures équivalentes, composées d'identificateurs et de variables. Le travail de ces modules est d'effectuer les unifications symboliques nécessaires pour déterminer toutes les structures communes cohérentes.

Les analyses effectuées sont de deux types :

• Celles à qui les informations véhiculées par

les deux îlots suffissent pour déterminer la cohérence de la jonction. Lorsque l'un des îlots peut être recouvert par les prédictions du deuxième par exemple. Ces analyses ne sont que des combinaisons de traitements formels extraits des modules de résolution de systèmes de contraintes.

• Celles qui ont besoin de connaissances supplémentaires pour conclure. C'est le cas par exemple quand les îlots ne présentent pas de prédictions (ni lexicales, ni syntaxiques). Le superviseur tente à l'aide des connaissances des experts de lier les deux sous-systèmes. Ainsi deux phonèmes "libres" (i.e. sans couverture lexicale ni syntaxique) pourront être joints si le superviseur trouve une couverture lexicale liant leurs deux systèmes de départ. Deux catégories syntaxiques sans prédiction devront être couvertes par une catégorie de niveau supérieur.

5. Implémentation

Notre système est actuellement implémenté en Prolog II+ (GIANNESINI, 1985), sur Macintosh II. Les fonctionnalités offertes par le langage Prolog (unifications et backtracking) sont utilisées par les modules de traitements formels pour la résolution des systèmes de contraintes et par le superviseur pour la gestion de l'algorithme.

Notre expert pour le Décodage Acoustico-Phonétique est le système développé au Laboratoire d'informatique d'Avignon (MELONI, 1991). Les connaissances syntaxiques et le module d'interprétation de la solution sont extraits de l'application ORBIS (interface d'interrogation en langue naturelle d'une base de données sur les planètes(COLMERAUER, 1982), les connaissances lexicales ont été complétées par des décompositions phonétiques étudiées sur un corpus test.

L'application (ORBIS) validant actuellement notre système est une application spécifique, à vocabulaire réduit (environ 150 mots amenant quelque 300 entrées phonétiques) pouvant traiter des phrases interrogatives avec relatives.

6. Conclusion

SYRAPAC est un système basé sur l'utilisation de connaissances explicites. Cette orientation a été prise dans la volonté de faire un produit adaptatif pouvant évoluer grâce à l'apport de nouvelles connaissances. Celles-ci pouvant provenir d'une meilleure expertise d'un domaine déjà exploité dans le système, mais aussi de l'inclusion d'un nouvel "expert". L'insertion d'une nouvelle source de connaissance est facilité par l'utilisation de traitements formels indépendants de l'interprétation des données. La structure du système n'exclut pas pour autant l'emploi de techniques stochastiques ou de réseaux neuronaux que nous pensons utiliser à terme pour faire intervenir les niveaux dits supérieurs dans la valuation des solutions. Le domaine d'application visé par SYRAPAC est l'interface vocale de petites et moyennes application en langage pseudo-naturel, possèdant un vocabulaire restreint.

BIBLIOGRAPHIE

COLMERAUER A., KITTREDGE R., ORBIS,9th international conference on computational Linguistics, COLING, 1982.

COLMERAUER A., Les grammaires de méthamorphoses, Communication homme-machine en langue naturelle avec déductions automatiques. Rapport final contrat Sesori n°73047 octobre 1976.

GIANNESINI F., KANOUI H., PASERO R., VAN CANEGHEM M., PROLOG, InterEdition 1985.

MELONI H., GILLES P., Décodage Acoustico-Phonétique ascendant, Traitement du Signal, Vol. 8 n°2, 1991.

WOODS W.A. Optimal search strategies for speech understanding control, Artifical Intelligence n°18 pp 295-326, 1982.

19es J.E.P. - Bruxelles, 19 au 22 mai 1992.

INTEGRATION DE LA DECOMPOSITION TEMPORELLE GENERALISEE DANS UN SYSTEME D'APPRENTISSAGE SYMBOLIQUE. APPLICATION A LA RECONNAISSANCE DES VOYELLES.

M.-J. CARATY & C. MONTACIE

LAFORIA - Université Paris 6, CNRS-URA 1095, 4, place Jussieu, 75252 Paris Cedex 5

Résumé

Charade est un système d'apprentissage symbolique qui engendre, à partir d'un ensemble d'exemples, un système de règles de production reflétant les régularités existant dans cet ensemble. Notre premier objectif est l'intégration, dans le processus d'apprentissage et de décision par règles, de la structuration temporelle a priori importante dans le signal vocal. Nous avons choisi une modélisation de l'évolution spectrale : la Décomposition Temporelle Généralisée (DTG). Notre second objectif est la reconnaissance des voyelles, il motive le choix de la distance par Appariement de Pics Spectraux (APS). Dans le processus de reconnaissance, trois représentations des voyelles sont testées, deux d'entre elles sont définies à partir de la DTG munie de APS. Les expériences utilisant le système Charade montrent que l'intégration de la structuration temporelle améliore de 7% le taux d'identification des voyelles (61 %). D'autres expériences utilisant un principe de décision par k-PPV au sens d'APS montrent que cette intégration n'apporte aucune amélioration par rapport au taux d'identification initial (i.e., sans DTG) de 81 %.

1. INTRODUCTION

Aussi divers que puissent être les systèmes de reconnaissance automatique, l'acquisition des connaissances fondée sur l'observation ou sur l'expérience est primordiale à leur performance. Une technique automatique de généralisation et d'apprentissage d'un système de règles de production, à partir d'un ensemble d'exemples et de contre-exemples, devrait permettre l'acquisition et la constitution d'une base de connaissances consistante.
Le système Charade [5], conçu pour détecter les régularités logiques ou statistiques existant dans un ensemble d'exemples, permet d'engendrer un système de règles de production reflétant ces régularités. Dans notre dernière étude [7] nous avons testé Charade, sur sa capacité à trouver des règles discriminant des macro-classes phonétiques. Nous avons amélioré le prétraitement pour prendre en compte l'évolution temporelle du signal de parole. Nous avons choisi pour cela un modèle original de l'évolution spectrale : la Décomposition Temporelle Généralisée [6]. Cette méthode étudie les déformations de la trajectoire spectrale pour localiser les principaux événements du signal de parole. La mesure de dissimilarité utilisée pour représenter la trajectoire spectrale est une mesure perceptive : l'Ajustement par Pics Spectraux [4].
La coopération du système Charade et de la technique de DTG munie de la mesure APS permet ainsi d'intégrer trois structurations a priori importantes dans les applications de Reconnaissance de la Parole : la structuration logique (i.e., les règles), la structuration temporelle (i.e., les événements) et la structuration fréquentielle (i.e., les pics spectraux).
L'évaluation de cette approche consiste en une série d'expériences sur la reconnaissance des voyelles.

2. STRUCTURATION FREQUENTIELLE : MESURE PAR APPARIEMENT DE PICS SPECTRAUX

Dans l'objectif de la reconnaissance des voyelles, notre choix de paramétrisation est motivé par l'importance bien connue des formants pour la caractérisation, la discrimination et la perception des voyelles. La représentation spectrale choisie est fondée, par analogie aux formants, sur les caractéristiques des maxima/pics spectraux. Dans cet espace de représentation, la mesure de distorsion inter-spectres utilisée est la mesure par Appariement de Pics Spectraux (APS). Cette mesure, fondée sur des critères de la perception sonore, a donné lieu pour gagner en robustesse à de nouveaux développements relativement à sa définition originelle [4].

2.1. Paramétrisation par pics spectraux

Un spectre à court-terme S est représenté par l'ensemble de ses pics spectraux $\{P_k\}_{(k=1,..,K)}$ caractérisés par leur fréquence centrale f_k (Hz), leur largeur de bande l_k (Hz) et leur amplitude a_k (dB) :

$$S = \{P_k(f_k, l_k, a_k)\}_{(k=1,..,K)}$$

Préalablement à son analyse par prédiction linéaire, le signal de parole est prétraité par une préemphase de +6dB/octave et par un fenêtrage de Hamming. Les fenêtres d'analyse sont de 25.6 ms, l'ordre de prédiction est fixé à 16.
Les pics (i.e., maxima locaux) sont directement détectés sur le spectre LPC logarithmique. La fréquence centrale et l'amplitude d'un pic P détecté à l'abscisse fréquentiel i sont estimées par interpolation parabolique passant par les trois points de l'enveloppe spectrale d'abscisses (i-1), i et (i+1). Par ordre de faisabilité, la largeur de bande de ce pic est -la largeur de bande à 3dB effectivement détectée sur l'enveloppe spectrale, -le double d'une demi-largeur de bande détectée, -l'interpolation parabolique passant par le point de l'enveloppe d'abscisse i et les deux points d'inflexion (i.e., gauche et droit) détectés au voisinage du pic. L'interpolation linéaire est systématiquement utilisée pour estimer les points caractéristiques cherchés (e.g., les points de chute à -3dB, les points d'inflexion).
Les diverses interpolations utilisées pour calculer les paramètres spectraux permettent de s'affranchir du seuil de quantification dû à la taille de la FFT employée. Il est a priori très important que la paramétrisation spectrale soit robuste et c'est ce que l'on peut constater a posteriori sur nos expériences d'identification des voyelles. De plus, la robustesse de la paramétrisation est indispensable pour utiliser la technique de Décomposition Temporelle Généralisée.

2.2. Mesure de distorsion APS

Si la mesure originelle par Appariement de Pics Spectraux [4] a montré de nombreuses qualités [8] [9], la recherche d'une meilleure robustesse a suscité de nouveaux développements. Ces développements ont consisté à définir, pour le calcul des distances inter-pics, les tables de pondération en fréquence, largeur de bande et amplitude sur l'échelle de fréquence considérée et à optimiser l'algorithme de calcul de la distance inter-spectres fondé sur l'appariement optimal des pics spectraux.

2.2.1. Distance locale inter-pics

La mesure de distance inter-pics $\delta(P_i, P_j)$, entre pics de spectres distincts, est introduite pour évaluer les distorsions spectrales locales. Elle est définie par :

$$\delta(P_i,P_j)=\omega_f(f_i).e_f(P_i,P_j)+\omega_l(f_i).e_l(P_i,P_j)+\omega_a(f_i).e_a(P_i,P_j)$$

où, $e_f(P_i,P_j) = \frac{|f_i-f_j|}{f_i+f_j}$; $e_l(P_i,P_j) = \frac{|f_i-f_j|}{f_i+f_j}$; $e_a(P_i,P_j) = |a_i-a_j|$
et $\omega_f(f_i)$, $\omega_l(f_i)$ et $\omega_a(f_i)$ sont des coefficients de pondération en fréquence, largeur de bande et amplitude.
Les tables de pondération ($\{\omega_f(f_i)\}$, $\{\omega_l(f_i)\}$, $\{\omega_a(f_i)\}$) sont calculées à partir des fonctions de répartition des quatre premiers formants sur l'échelle de fréquence considérée ($\{F_1(f_i)\}$, $\{F_2(f_i)\}$, $\{F_3(f_i)\}$, $\{F_4(f_i)\}$) [3] et des valeurs-type $\{\Omega_f(F_i), \Omega_l(F_i), \Omega_a(F_i)\}_{(i=1,..,4)}$ [4].

2.2.1. Distance globale inter-spectres

La distance globale APS est calculée à partir de l'appariement optimal des spectres considérés. Soient S_R et S_T deux spectres distincts, l'appariement d'un pic P_i^R de S_R à un pic P_j^T de S_T est évalué par la distance inter-pics $\delta(P_i^R, P_j^T)$ et il est défini optimal lorsque :

$$P_j^T = \text{Argmin}_{\{P^T \in S^T\}} \{\delta(P_i^R, P^T)\}$$

Soit Δ la matrice des distances inter-points des spectres considérés :

$$\Delta = \{\Delta_{ij} = \delta(P_i^R, P_j^T)\}_{(i=1,..,I\ ;\ j=1,..,J)}$$

La distance globale inter-spectres $D_{APS}(S^R, S^T)$ est la moyenne des appariements optimaux distincts obtenus en balayant les lignes et les colonnes de la matrice Δ.

3. STRUCTURATION TEMPORELLE : DECOMPOSITION TEMPORELLE GENERALISEE

La Décomposition Temporelle classique (DT) [1] modélise l'évolution spectrale d'un segment de parole. Elle utilise un modèle d'interpolation linéaire et des techniques d'analyse ayant comme critère la minimisation d'une erreur quadratique. L'opérateur de reconnaissance, si l'on veut rester rigoureux, doit être fondé sur la distance euclidienne. Cette distance, si elle est simple à mettre en oeuvre (d'où son succès), n'est pas la plus performante des mesures de dissimilarité spectrale. La Décomposition Temporelle Généralisée [6], pour sa part, utilise un modèle d'interpolation linéaire elle-même généralisée à une mesure de dissimilarité D quelconque.

3.1. Modélisation par DT

La Décomposition Temporelle [1] décrit l'évolution spectrale d'un segment de parole, représenté par les vecteurs spectraux $\{y_n\}_{(n=1,..,N)}$, par un modèle d'interpolation linéaire :

$$\hat{y}_n = \sum_{i=1}^{q} g_i\, \phi_i(n)$$

L'estimation $\hat{y}_n$ du vecteur spectral y_n est la combinaison linéaire d'un nombre limité de vecteurs spectraux $\{g_i\}_{(i=1,..,q)}$, appelés "cibles" spectrales. A chaque cible g_i est associée une fonction d'interpolation à support compact ϕ_i. Les cibles $\{g_i\}$ expriment le contenu spectral du signal, la fonction d'interpolation $\phi_i(n)$ mesure l'influence de la $i^{ème}$ cible pour l'estimation du $n^{ème}$ vecteur spectral y_n. Le but de la Décomposition Temporelle est d'estimer à la fois les fonctions d'interpolation $\{\phi_i\}$ et les cibles spectrales $\{g_i\}$ (cf. Figure 2). La seule contrainte imposée a priori concerne la forme des fonctions ϕ_i : elles doivent être positives et compactes dans le temps, c'est-à-dire avoir des valeurs non-nulles sur un intervalle de temps fini.

3.2. Modélisation par DTG

Pour généraliser la Décomposition Temporelle [6], nous avons proposé une approximation du modèle d'interpolation linéaire généralisée en utilisant l'analyse factorielle. Il s'agit de projeter une partie de la trajectoire spectrale dans un nouvel espace muni de la distance euclidienne, et de rechercher la fonction d'interpolation représentant ce segment de signal.

Soit la trajectoire spectrale représentée par les vecteurs $\{y_n\}_{(n=1,..,N)}$, d'un espace métrique quelconque muni d'une distance D. Pour estimer la fonction d'interpolation il faut calculer les n premiers vecteurs propres de la matrice S de taille NxN dont les éléments $\{S_{ij}\}_{(i,j=1,..,N)}$ sont calculés par la formule :

$$S_{ij} = -\sum_{k,l=1}^{N} D(y_k, y_l)^2 / 2N^2 - D(y_i, y_j)^2 / 2 + (\sum_{k=1}^{N} D(y_i, y_k)^2 + \sum_{k=1}^{N} D(y_k, y_j)^2) / 2N$$

Les n premiers vecteurs propres de S doivent représenter au moins 95 % de l'inertie de la trajectoire (i.e., 95 % de la trace de S). La fonction d'interpolation choisie est une combinaison linéaire des n premiers vecteurs propres. Cette combinaison linéaire doit maximiser la corrélation entre la fonction d'interpolation calculée et une fonction caractéristique rectangulaire.

4. STRUCTURATION LOGIQUE : APPRENTISSAGE SYMBOLIQUE PAR CHARADE

Le système Charade [5], conçu pour détecter les régularités logiques ou statistiques existant dans un ensemble d'exemples, permet d'engendrer un système de règles de production reflétant ces régularités. La technique d'apprentissage est fondée sur l'utilisation d'une structure de données particulière, le cube de Hilbert, et sur une exploration "intelligente" de l'espace de description pour la génération des règles.

4.1. Technique d'apprentissage de Charade

Le système CHARADE (Cubes de Hilbert Appliqués à la Représentation et à l'Apprentissage à partir de Descriptions d'Exemples), à partir -d'un langage de description, -d'un ensemble d'axiomes reflétant la sémantique du langage et -d'un ensemble d'exemples exprimés dans ce langage, engendre un système de règles de production consistant.

4.1.1. Langage de description des exemples

Le langage de description des exemples est inclus dans le Langage de la Logique des Propositions. La description d(E) d'un exemple E est une conjonction logique de descripteurs :

$$d(E) = d_1 \wedge d_2 \wedge \ldots \wedge d_D$$

Chaque descripteur d_i est originellement une proposition atomique ou la négation d'une proposition atomique. Cette description des exemples permet d'évaluer rapidement les opérations élémentaires dans le domaine de l'apprentissage. Ainsi, la généralisation et la discrimination s'évaluent par de simples conjonctions de descripteurs. Par exemple, la généralisation minimum de deux exemples est la conjonction des descripteurs communs aux deux exemples. Le langage de description peut être étendu à l'utilisation de descripteurs multi-valués.

4.1.2. Espace de représentation des exemples

L'espace de représentation de l'ensemble d'apprentissage (e.g., un ensemble de n exemples) est un cube de Hilbert : un hypercube unitaire dans un espace à n dimensions. Dans cet espace de représentation :

— chaque axe est associé à un exemple,

— chaque sommet correspond à un sous-ensemble d'exemples caractérisé par sa généralisation minimum,

— l'appartenance d'un exemple à un sommet est définie par la projection du sommet sur l'axe correspondant,

— les arètes du cube sont considérées comme des relations d'héritage orientées vers l'origine. Ces relations d'héritage permettent une optimisation de la représentation : un sommet hérite de tous les sommets au-dessus de lui dans la hiérarchie du cube.

4.1.3. Principe de génération des règles logiques

Le principe d'induction est le suivant : soient deux descripteurs d_1 et d_2, si tous les exemples de l'ensemble d'apprentissage qui contiennent d_1 contiennent également d_2 dans leur description, alors $d_1 \Rightarrow d_2$.

Pour introduire le principe de génération, considérons les deux espaces de représentation suivants :

— Le Cube de Hilbert des Exemples (C_{Ex}) qui représente l'ensemble des parties de l'ensemble d'apprentissage. C_{Ex} est ordonné par l'inclusion ensembliste.

— Le Cube de Hilbert des Descripteurs (C_{Des}) qui représente l'ensemble des conjonctions de descripteurs. C_{Des} est ordonné par l'implication logique.

Le principe de génération des règles consiste alors à définir le lien entre ces deux relations d'ordre par une exploration du cube des descripteurs.

Pour la génération d'une règle potentielle, quatre fonctions sont définies.

Deux fonctions δ et γ sont définies : δ (resp. γ) associe à chaque sommet de C_{Ex} (resp. C_{Des}) un sommet de C_{Des} (resp. C_{Ex}).

— $\delta : C_{Ex} \longrightarrow C_{Des}$. Pour chaque sommet de C_{Ex} (i.e., un ensemble d'exemples), δ associe le sommet de C_{Des} correspondant à la généralisation minimum de l'ensemble considéré.

— $\gamma : C_{Des} \longrightarrow C_{Ex}$. A chaque sommet de C_{Des} (i.e., une conjonction de descripteurs), γ associe le sommet de C_{Ex} correspondant à l'ensemble de tous les exemples pour lesquels la conjonction de descripteurs apparaît.

L'application $\beta = \delta \circ \gamma : C_{Des} \longrightarrow C_{Des}$, a la propriété de faire apparaître toutes les relations logiques présentes dans l'ensemble d'apprentissage.

— ω et τ sont définies pour éliminer les redondances des relations logiques relativement aux relations d'héritage et à la transitivité de l'implication.

Avec les définitions des fonctions données précédemment, pour chaque sommet S_{Des} du cube des descripteurs il est possible de générer une règle du type :

$$S_{Des} => \tau(\omega(\beta(S_{Des})))$$

Pour la génération du système complet des règles de production, une exploration exhaustive du cube des descripteurs est exclue. Pour la faisabilité de la technique de génération des règles, plusieurs théorèmes de limitation de l'exploration du cube des descripteurs sont utilisés.

4.1.4. Théorèmes de limitation

Deux théorèmes fondamentaux déterminent le critère de nullité des sommets du cube des descripteurs (i.e., les sommets pour lesquels $\tau \circ \omega \circ \beta$ est vide). Le but est d'éliminer les sommets inutiles (i.e., les sommets qui n'engendrerons pas de nouvelles règles).
D'autres théorèmes sont introduits comme des contraintes au système des règles de production. Ces contraintes sont principalement liées aux propriétés désirées du système des règles engendré. Par exemple, la pertinence des descripteurs, le fractionnement des classes, la structuration des règles, la couverture des exemples, le facteur de bruit, la condition terminale des règles, etc... Le système Charade admet une quinzaine de paramètres à ajuster.

4.2. Paramétrisation par exemples

Un spectre à court-terme est représenté par l'ensemble de ses pics spectraux (cf. §2.1.). 30 descripteurs binaires $\{d_k\}_{(k=1,..,30)}$ sont calculés à partir de cette représentation. Soient m et v les deux valeurs de descripteur autorisées : d_k=m (resp. d_k=v) correspond à la présence d'une masse (resp. vallée) spectrale dans la bande de fréquence B_k (cf. Figure 1). Les bandes de fréquence B_k sont contigües et ont une largeur de 63 Mels.

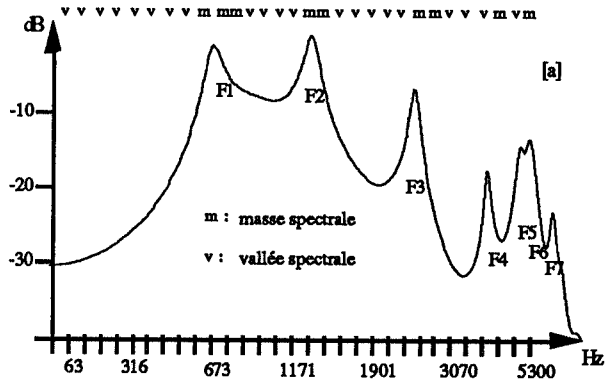


Figure 1. Illustration de la description d'un exemple

5. EXPERIMENTATIONS

Nous limitons notre application du système Charade à l'identification des voyelles orales/nasales dans un signal de parole continue. L'apprentissage des entités vocaliques engendre un système de règles pour la classification des voyelles : chaque règle concluant sur le code phonétique d'une voyelle. Le principe d'identification d'une entité-test est très simple et rapide, il est fondé sur le comptage des déclenchements de règles apprises sur l'entité-test.
Les entités considérées posent le problème de leur représentation. En effet, ces entités sont des continuums de signal : des réalisations phonétiques de voyelles contenues dans le signal vocal. Plusieurs représentations des entités-voyelles sont testées, elles prennent en compte localement ou globalement la structuration du signal de parole. Pour situer les résultats obtenus, nous choisissons pour points de comparaison, des expériences de reconnaissance des formes non-paramétrique.

5.1. Base de données

Les expériences ont été effectuées sur les corpus SYL_y-Acoustique de la base de données BDSONS-CNRS-GRECO enregistrés par un locuteur masculin (BP). Ces corpus SYL_y, numérotés (y) de 1 à 12, permettent l'étude des 192 diphones $[C_nV_y]$ dénombrés pour les voyelles orales ou nasales $\{[V_y]\}_{(y=1,..,12)}$ et les consonnes $\{[C_n]\}_{(n=1,..,16)}$ du français. Un corpus SYL_y est constitué de 16 phrases $\{P_n\}_{(n=1,..,16)}$, une phrase P_n contenant plusieurs occurrences du diphone $[C_nV_y]$.
La base de données comprend plus de 4000 phonèmes dont 1980 voyelles orales/nasales. Les signaux de parole ont été étiquetés manuellement sur le principe de l'étiquetage large [2]. Lors de cet étiquetage, certaines distinctions phonétiques n'ont pas été faites (e.g., {[œ], [ə], [ø]}, {[ɛ], [œ̃]}). Les 1980 voyelles se répartissent ainsi en 12 classes vocaliques (i.e., [a], [i], [e], [ɛ], [y], [ø], [u], [o], [ɔ], [ɑ̃], [ɛ̃], [ɔ̃]).

5.2. Représentation des voyelles

De la plus élémentaire à la plus "sophistiquée", les représentations testées pour les voyelles sont les suivantes :
— Représentation $Fen_{Etiq\ Large}$: la voyelle est représentée par la fenêtre de signal centrée sur la localisation temporelle de l'étiquette posée lors de l'étiquetage large. Il s'agit d'une représentation locale qui correspond au "centre" de la réalisation du son (i.e., généralement le point d'énergie maximale).
— Représentation $Fen_{g(DTG,\ APS)}$: la voyelle est représentée par la fenêtre centrée sur le centre de gravité de la fonction d'interpolation calculée par la DTG munie de la distance APS. Il s'agit d'une représentation locale déduite de la modélisation de l'évolution spectrale par DTG et qui peut être interprétrée comme une caractéristique spectrale du son considéré.

— Représentation $Seg_{\phi(DTG,APS)}$: la voyelle est représentée par le segment défini par le support temporel de la fonction d'interpolation calculée par la DTG munie de la distance APS. Il s'agit d'une représentation globale également déduite de la DTG et qui reflète la contribution segmentale de l'événement acoustique considéré.

5.3. Apprentissage et principe de décision par règles

Pour chaque voyelle, l'ensemble d'apprentissage est constitué de la représentation de la première occurrence de la voyelle [V] en contexte $[C_nV]_{(n=1,..,16)}$, soit 16 formes-référence par voyelle.
Le système Charade est paramétré de telle sorte que chaque règle engendrée conclut sur une condition terminale : le code phonétique d'une voyelle apprise. Charade ne permet pas l'apprentissage d'un continuum. Par conséquent, parmi les représentations des sons sélectionnées seules les représentations locales pourront être considérées à l'apprentissage. Ainsi, deux systèmes de règles effectivement engendrés par Charade seront testés. Un troisième système de règles, simple concaténation des deux premiers, sera également testé.

L'ensemble de test, disjoint de l'ensemble d'apprentissage, est constitué des 1789 formes restantes de la base de données. Les trois représentations sélectionnées sont testées pour les formes-test.
Le principe de décision par règles consiste à prendre une décision suivant le nombre, par voyelle, de déclenchements des règles sur la forme-test considérée. A chaque voyelle est associé un compteur du nombre de déclenchements des règles concluant sur cette voyelle. A l'identification d'une forme-test, pour chaque règle du système de règles engendré dont les prémisses sont vérifiées par la forme-test, on incrémente le compteur associé à sa condition terminale. Si l'un des compteurs est le maximum maximorum, le résultat est une Identification ou une Substitution, dans le cas contraire le résultat est un Rejet.

Les résultats de reconnaissance (cf. Tableaux 1) sont donnés pour chacun des trois systèmes de règles sélectionnés et pour deux représentations des formes-test : l'une locale (notée I) et l'autre globale (notée II).

Charade $App_{Etiq\ Large}$	Ident. %	Subst. %	Rejet. %
I $Test_{Etiq\ Large}$	53,5%	28,5%	17,9%
II $Test_{\phi(DTG,APS)}$	59,4%	34,3%	6,4%

Charade $App_{g(DTG,APS)}$	Ident. %	Subst. %	Rejet. %
I $Test_{g(DTG,APS)}$	49,2%	32,7%	18,1%
II $Test_{\phi(DTG,APS)}$	54,8%	39,5%	5,6%

Charade $App_{Etiq\ Large}$ & $App_{g(DTG,APS)}$	Ident. %	Subst. %	Rejet. %
I $Test_{Etiq\ Large}$	60,7%	31,3%	8,0%
II $Test_{\phi(DTG,APS)}$	61,4%	35,8%	2,9%

Tableaux 1. Résultats de reconnaissance du processus d'apprentissage et de décision par règles

Quelque soit le système de règles considéré, les résultats montrent que les taux d'identification sont meilleurs pour une identification des segments-test.
Les centres de gravité des fonctions d'interpolation de la DTG ne se révèlent pas une information stratégique pour l'identification.
L'enrichissement d'un système de règles se révèle pertinent, les taux d'identification augmentent d'environ 7% quelque soit la représentation des formes-test.

5.4. Principe de décision par k-Plus Proches Voisins

Cette expérience ne permet en rien une comparaison directe avec les résultats obtenus par Charade. Elle permet simplement de situer les expériences précédentes relativement à l'intégration de la structuration temporelle. L'expérience est un test-fermé où chaque entité de la base de données est comparée aux autres. La liste des étiquettes des entités les plus proches au sens de la distance considérée (i.e., la distance APS pour les représentations locales et la comparaison dynamique DYN pour la représentation globale) est engendrée. Le principe d'identification d'une entité est la règle du k-Plus Proche Voisin (k-PPV) avec vote maximum maximorum (cf. §5.3.). Selon ce principe, les résultats se répartissent en trois catégories : Identification, Substitution et Rejet.

Les résultats de reconnaissance (cf. Tableaux 2) sont donnés pour chacune des représentations sélectionnées et pour deux valeurs de k : k=1 et $k=k_{opt}$ ($k_{opt}\leq 20$) qui donne le meilleur taux d'identification.

Test-Fermé $Fen_{Etiq\ Large}$	Ident. %	Subst. %	Rejet. %
PPV_{APS}	76,6%	23,4%	—
13-PPV_{APS}	81,5%	16,6%	1,9%

Test-Fermé $Fen_{g(DTG,APS)}$	Ident. %	Subst. %	Rejet. %
PPV_{APS}	74,4%	25,6%	—
13-PPV_{APS}	77,8%	20,1%	2,1%

Test-Fermé $Seg_{\phi(DTG,APS)}$	Ident. %	Subst. %	Rejet. %
PPV_{DYN}	74,6%	25,4%	—
11-PPV_{DYN}	77,0%	19,8%	3,2%

Tableaux 2. Résultats de reconnaissance du principe de décision des k-PPV

Nos expériences montrent que des informations a priori stratégiques, telles celles obtenues par la DTG (i.e., localisation ou segmentation), ne se sont pas avérées des informations utiles pour l'identification. Néanmoins, aucune conclusion ne peut apportée quant aux défauts de la modélisation ou de la distance utilisée.

6. CONCLUSION

Nous avons présenté une approche pour prendre en compte l'évolution temporelle du signal de parole dans un processus d'apprentissage symbolique et de reconnaissance par règles. Les résultats de nos expériences avec le système Charade ont montré que l'intégration de la structuration temporelle améliore sensiblement le taux d'identification des voyelles. Le taux de reconnaissance obtenu par un système de décision par k-PPV est significativement plus élevé. Néanmoins, la description des exemples dans le système d'apprentissage symbolique est particulièrement simple et le temps de reconnaissance est sans commune mesure. L'amélioration du langage de description et une description des exemples plus pertinente (e.g., forme de la fonction d'interpolation de la DTG) devraient permettre de réduire les différences de performances du système testé relativement à des systèmes de décision plus élaborés.

REFERENCES

[1] B.S. Atal, "Efficient Coding of LPC Parameters by Temporal Decomposition", IEEE ICASSP, pp. 81-84, 1983.

[2] L.-J. Boë & L. Miclet, "Manuel d'étiquetage large. GRECO n°39 Communication Parlée", Commis-sion Etiquetage BDSON & EUROM, 1988.

[3] L.-J. Boë, P. Perrier & G. Bailly, "The Geometric Vocal Tract Variables Controlled for Vowel Production : Proposal for Constraining Acoustic-to-Articulatory Inversion", Journal of Phonetics, n°20, pp. 27-37, 1992.

[4] M.-J. Caraty, "Contribution au décodage acoustico-phonétique : études de distances inter-spectres et reconnaissance de cycles vocaliques", Thèse de l'Université Paris 6, 1987.

[5] J.-G. Ganascia, "Learning with Hilbert Cubes", 2nd European Working session on Machine Learning, Sigma Press, pp. 158-171. 1986.

[6] C. Montacié, "Décodage acoustico-phonétique : apport de la décomposition temporelle généralisée et de transformations spectrales non-linéaires. Application à la reconnaissance des mots épelés en continu.". Thèse de l'ENST, 1991.

[7] C. Montacié, M.-J. Caraty & X. Rodet, "Experiments in the Use of an Automatic Learning System for Acoustic-Phonetic Decoding", ICSLP-90, pp. 357-390, 1990.

[8] H. Ye, M.-J. Caraty, L.-J. Boë & D. Tuffelli, "Distances interspectrales et macro-sensibilité des voyelles focales", 17èmes JEP, pp. 20-25, 1988.

[9] H. Ye, M.-J. Caraty, L.-J. Boë & D. Tuffelli, "Structural Phonetic Evaluation of Dissimilarities Functions Used in Speech Recognition, Eurospeech, pp. 404-407, 1989.

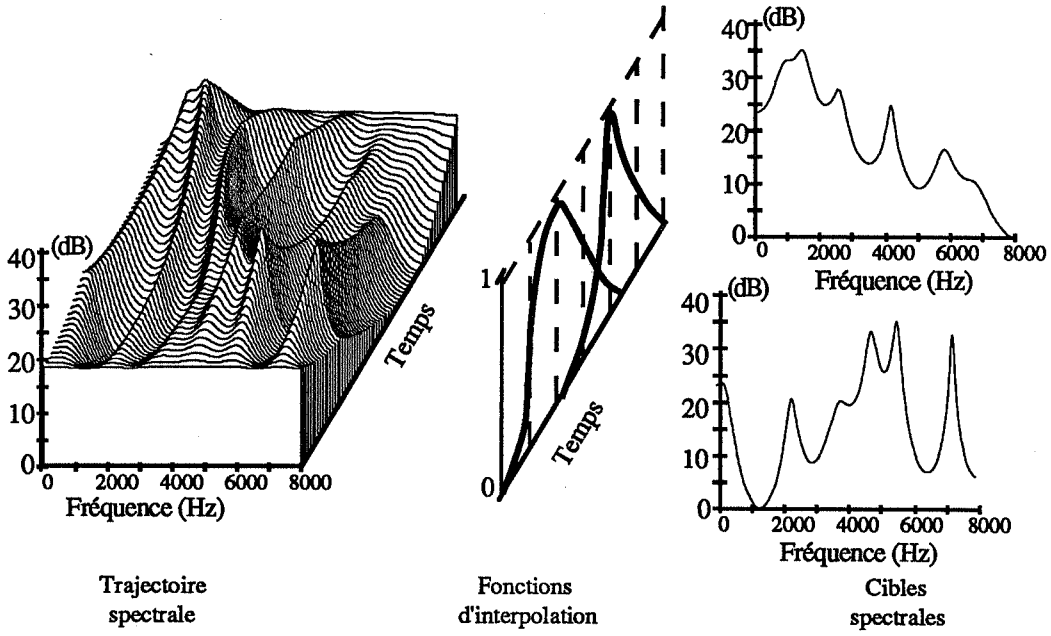


Figure 2. Exemple de Décomposition Temporelle sur le segment de parole [ia]

19es J.E.P. - Bruxelles, 19 au 22 mai 1992.

UNE MÉTHODE CENTISECONDE POUR LA RECONNAISSANCE D'UN GRAND VOCABULAIRE DE MOTS ISOLÉS

Mohamed NAIT-LAHCEN[2], *Gilles ADDA*[1], *Stéphane BORNERAND*[2]

1 LIMSI-CNRS, B.P. 130, F-91403, ORSAY Cedex, FRANCE

2 BULL S.A, 7,Rue Ampère 91343, MASSY Cedex, FRANCE

Résumé

Le travail décrit dans cet article entre dans le cadre du projet ESPRIT-POLYGLOT1 *N°*2104 . L'un des buts de ce projet est de produire un système de reconnaissance multi-langue de mots isolés, utilisant le même matériel et logiciel. De ce fait, le système de reconnaissance italien réalisé par OLIVETTI, a été choisi pour être adapté à 6 autres langues européennes. Les prototypes seront la seule information dépendante du locuteur dans tout le système de reconnaissance: un prototype représente 10*ms* de parole, et il est extrait des phonèmes stables. Le module de construction de chaînes phonétiques propose une "pseudo" chaîne phonétique, utilisée par le module de présélection afin de tirer du vocabulaire un certain nombre de candidats. Ces candidats seront traités par le module d'analyse phonétique fine, afin d'en extraire les *N* meilleurs.
Dans cet article, nous donnerons les résultats d'une série d'expériences en présélection, menée sur 10 locuteurs (5 hommes et 5 femmes) et des dictionnaires de 1000, 2000 et 8000 mots.

1 INTRODUCTION

Le travail décrit dans cet article entre dans le cadre du projet ESPRIT-POLYGLOT1 *N°*2104 . L'un des buts de ce projet est de produire un système de reconnaissance multi-langue, utilisant le même matériel et logiciel. De ce fait, le système de reconnaissance de grands vocabulaires (en mots isolés), réalisé pour la langue italienne (Billi et. al.[1], 1989), a été choisi pour être adapté à 6 langues européennes; la langue anglaise, allemande, hollandaise (Drexler et. al.[2], 1991), espagnole, greque, et enfin la langue française.
Le système de reconnaissance est conçu de façon modulaire (voir figure 1). Chaque module nécessite un certain nombre d'outils, tous ces outils étant préparés lors d'une phase d'apprentissage à partir de plusieurs statistiques.
A partir de 10 locuteurs, une base de données de parole segmentée et étiquetée phonémiquement de façon manuelle, a servi pour calculer des statistiques de durée des phonèmes dans différents contextes. Un corpus de texte, tiré à partir du journal "LE MONDE", contenant environ 5000000 de mots, a servi pour le calcul des statistiques concernant la fréquence des phonèmes et les paires de phonèmes.
Le signal analogique issu d'un microphone, est traité d'abord par le module d'analyse et traitement du signal. Les autres modules qui interviennent dans le système de reconnaissance sont :

- Le module de détection des prototypes : cette détection est faite par comparaison de chaque trame du signal analysé (ce qui donne le caractère centiseconde au système de reconnaissance) à une liste de prototypes. Ces derniers sont extraits lors de la phase d'apprentissage, et seront par la suite la seule information dépendante du locuteur dans tout le système de reconnaissance.
- Le module de construction de chaînes "pseudo" phonétiques : ces chaînes sont ainsi appelées car elles sont obtenues par concaténation et regroupement des prototypes précédemment détectés. Les vraies chaînes phonétiques sont listées dans le lexique et seront désignées par *modèles phonétiques* dans la suite de cet article, alors que ces "pseudo" chaînes seront désignées par *chaînes phonétiques.*
- Le module de présélection : il sert à extraire d'un grand vocabulaire un certain nombre de candidats. Ceci a pour effet de réduire la taille du vocabulaire dynamique. Cette réduction peut atteindre 90% de la taille totale du vocabulaire dynamique tout en gardant de bonnes performances (Billi [3], 1986).
- Le module d'analyse phonétique fine : il utilise des statistiques de durée de phonèmes

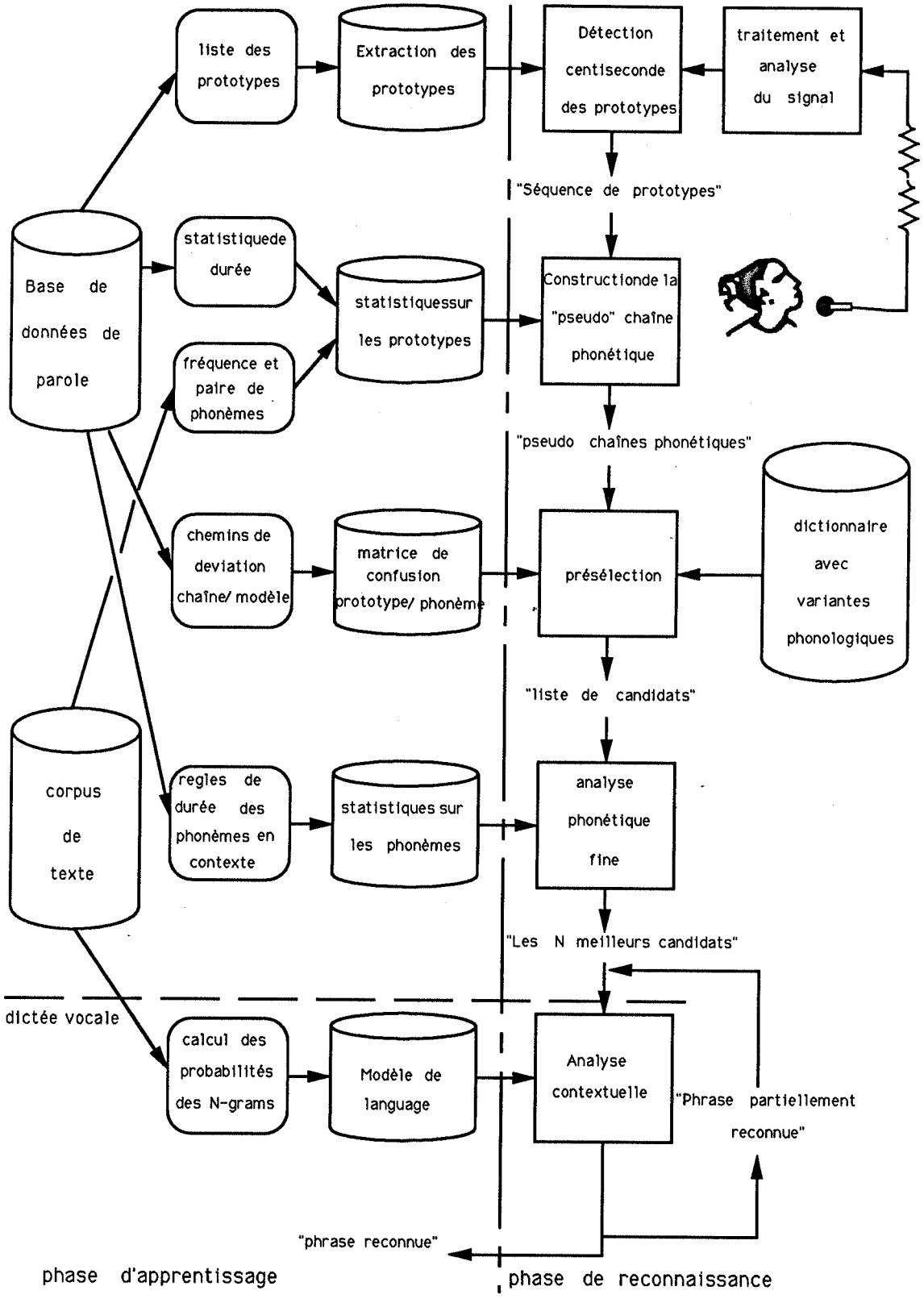


Figure 1: Schéma fonctionnel du système Polyglot

en contexte, des statistiques sur l'énergie des phonèmes, et des statistiques de distances entre les prototypes et tous les phonèmes pour extraire, à partir des candidats délivrés par le module de présélection, les N meilleurs candidats.

- Le module d'analyse contextuelle : il permet de faire de la dictée vocale en utilisant un modèle de langage probabiliste obtenu à partir d'un corpus de texte.

Nous avons concentré notre étude sur le module de présélection car les performances du système de reconnaissance dépend étroitement de ce module. Nous donnerons les résultats d'une série d'expériences de présélection menée sur 10 locuteurs (5 hommes et 5 femmes), ayant prononcés chacun 1000 mots.

2 ANALYSE DU SIGNAL

Toutes les $10ms$, une préaccentuation et une fenêtre de Hamming de longueur $30ms$ sont appliquées sur le signal de parole. Cette portion du signal est analysée par prédiction linéaire, délivrant 20 coefficients d'autocorrélation. A partir de ces derniers, l'algorithme Leroux-Guegen [4] permet de calculer les coefficients de réflexion qui, à leur tour, permettent de déduire les coefficients cepstraux grâce à la méthode Levinson-Durbin [5].
En définitive, l'analyseur acoustique délivre 20 coefficients d'autocorrélation, 20 coefficients cepstraux et deux valeurs d'énergie (avec et sans préaccentuation), toutes les $10ms$. Ces deux jeux de coefficients permettent de calculer la distance WLR qui permet de comparer les trames avec les différents prototypes. L'énergie avec préaccentuation sert à la détection du début et fin du mot.

3 EXTRACTION DES PROTOTYPES

Un prototype est un vecteur à 20 paramètres, qui peuvent être soit les 20 coefficients d'autocorrélation, soit les 20 coefficients cepstraux. Le prototype est défini afin de représenter la trame la plus "caractéristique" d'un phonème. Cette trame est obtenue par moyennage de toutes les trames appartenant à la zone la plus stable du phonème (le moyennage est effectué sur les coefficients de réflexion). La distance inter-trame, calculée le long du signal, est utilisée afin de déduire une fonction de stabilité des trames, faisant correspondre le maximum de stabilité au minimum de distance inter-trame.
Par sa définition, un prototype ne peut représenter qu'un phonème présentant une zone de stabilité. De ce fait, les plosives et les semi-voyelles, dont le signal présente peu de stabilité, ne seront représentées par aucun prototype au niveau de la présélection; par contre, lors de l'analyse phonétique fine, tous les phonèmes, sans aucune exception, seront représentés par au moins un prototype. Pour les phonèmes présentant une zone de stabilité, nous avons testé deux listes de prototypes: une première liste contenant 17 prototypes obtenus en appliquant l'algorithme des k-moyennes sur l'ensemble des phonèmes pour déterminer un certain nombre de classes. L'autre liste contient 24 prototypes correspondant aux phonèmes stables de la langue française.
La seule information dépendant du locuteur dans tout le système de reconnaissance étant le prototype, il est plus intéressant d'avoir une liste de prototypes plus complète. Nous avons donc opté pour la deuxième liste. Afin de tenir compte des allophones, un prototype est obtenu par moyennage de plusieurs prototypes correspondant au même phonème dans différents contextes.

4 CONSTRUCTION DE LA CHAÎNE PHONÉTIQUE

Cet étape consiste à transformer le signal d'un énoncé en une suite de prototypes. Ces derniers doivent identifier approximativement les phonèmes stables du mot. Cette suite de prototypes constitue une chaîne phonétique, qui servira, lors de la présélection, à extraire du vocabulaire un certain nombre de candidats, qui seront traités lors d'une analyse phonétique plus fine.
Chaque trame du mot prononcé est comparée à la totalité des prototypes. La comparaison est faite grâce à la distance WLR ("Weighted Likelihood Ratio" Shikano, [6], 1981) dont la formule est la suivante:

$$d_{WLR}(x,y) = \frac{1}{2}\sum_{1}^{N}\left(\frac{r_i}{r_0} - \frac{r'_i}{r'_0}\right)(c_i - c'_i)$$

où

r et r' représentent les coefficients d'autocorrélation. c et c' représentent les coefficients cepstraux.

Expérimentalement, il a été montré que pour un nombre de paramètres supérieur à 16, le domaine de sommation de cette distance peut se réduire au nombre de paramètres au lieu d'être infini. Dans notre cas, comme nous utilisons 20 paramètres, nous n'avons jamais rencontré de distance négative.
Les prototypes sont alors ordonnés dans un tableau en fonction de leur distance respective avec la trame courante du mot, constituant ainsi, le long du mot, un treillis de prototypes. La programmation dynamique est un algorithme efficace pour trouver le chemin optimal sur le treillis de prototypes.
Une fonction de coût est définie afin de comparer les différents chemins. Le coût de chaque chemin est obtenu en additionnant le coût local en chacun de ses noeuds. Ce coût local est calculé de façon pseudo-probabiliste:

$$cout_{local} = -\log_{10}[p(ph/a,b,c,d)]$$

où

ph : Un phonème stable.
a : La distance WLR affectée aux prototypes.
b : Statistique de durée.
c : Statistique sur la fréquence des phonèmes.
d : Statistique sur les paires de phonèmes (régles phonotactiques).

En utilisant le théorème de Bayes l'égalité devient :

$$cout_{local} = -\log_{10}[\frac{p(a,b,c,d/ph)}{p(a,b,c,d)} * p(ph)]$$

Si on suppose que les statistiques sont indépendantes:

$$cout_{local} = -\log_{10}\frac{p(a/ph)}{p(a)} - \log_{10}\frac{p(b/ph)}{p(b)} - \log_{10}\frac{p(c/ph)}{p(c)} - \log_{10}\frac{p(d/ph)}{p(d)} - \log_{10}[p(ph)]$$

En d'autres termes:

$cout_{local}$ = score de distance + score de durée scores de fréquence + ...

Tout ces scores proviennent des statistiques faites lors de la phase d'apprentissage. Le coût total est obtenu en additionnant les coûts locaux le long du chemin optimal:

$$cout_{total} = \sum_i cout_{local}(i)$$

Le chemin optimal est celui qui minimise la fonction de coût. Les noeuds du chemin optimal délivrent une suite de prototypes (voir figure 2). En évitant les répétitions adjacentes d'un même prototype, nous obtenons la chaîne phonétique.

5 LA PRÉSÉLECTION

Il s'agit, dans cette étape, de comparer la chaîne phonétique à tous les modèles phonétiques des mots du dictionnaire, afin de déterminer un sous-ensemble du vocabulaire utilisé dans la suite du traitement. Cette comparaison est faite par programmation dynamique, en alignant la chaîne phonétique à chacun des modèles du dictionnaire.
Les contraintes locales du chemin de déviation sont:

- $C_S(x_i, y_j)$: Probabilité de substitution du symbole x_i (prototype) de la chaîne phonétique, avec le phonème y_j du modèle.
- $C_I(x_i)$: Probabilité d'insertion du symbole x_i.
- $C_D(y_j)$: Probabilité d'élision du phonème y_j.

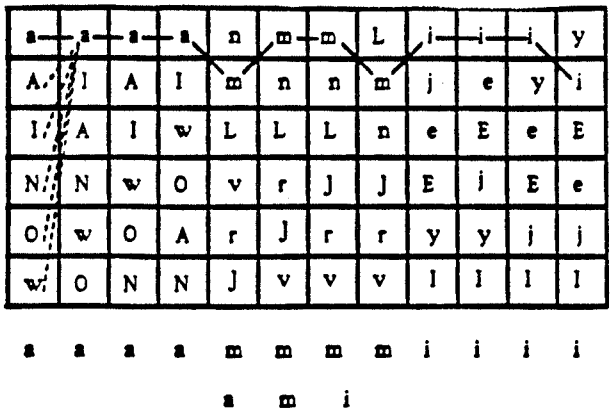


Figure 2: Programmation dynamique sur le treillis de prototypes

Afin de simplifier l'estimation de ces probabilités, la chaîne phonétique étant généralement plus longue que le modèle correspondant à cause des zones de transition, nous pouvons supposer que $C_I(x_i) = C_S(x_i, y_j)$ si (x_{i-1}, y_j) est sur le chemin optimal. Utilisant les deux tableaux $C_S(x_i, y_j)$ et $C_D(y_j)$, une simple programmation dynamique est effectuée de la manière suivante:

condition initiale: $g(c(1)) = d(c(1))$
équation de récursion:
$g(c(k)) = min_{c(k-1)}[g(c(k-1)) + d(c(k))]$
où $d(c(k))$ est définie par :

$$\mathrm{d(c(k))} = \begin{cases} C_D(y_{j(k)}) & \text{si } i(k-1) = i(k) \\ C_I(x_{i(k)}) & \text{si } j(k-1) = j(k) \\ C_S(x_{i(k)}, y_{j(k)}) & \begin{cases} \text{si } i(k) = j(k-1) + 1 \\ \text{et } j(k) = j(k-1) + 1 \end{cases} \end{cases}$$

Les N plus faibles distances délivrent les N candidats reconnus par l'étage de présélection.

6 LA MATRICE DE CONFUSION

Les contraintes locales utilisées par la programmation dynamique lors de la présélection $C_S(x_i, y_j)$ et $C_D(y_j)$ (voir figure 3), correspondent respectivement à la matrice de confusion et d'élision. Calculer ces deux matrices, revient à estimer les probabilités de substitution et d'élision. Plusieurs méthodes d'estimation de ces fonctions de coût peuvent être utilisées; nous avons adopté la méthode du maximum de vraisemblance.

$$M^* = argmax\ p(O/M)$$

où O est une chaîne phonétique délivrée par le module de construction de la chaîne phonétique, et M est le modèle qui devrait générer cette chaîne phonétique par une suite de substitution et d'élision.

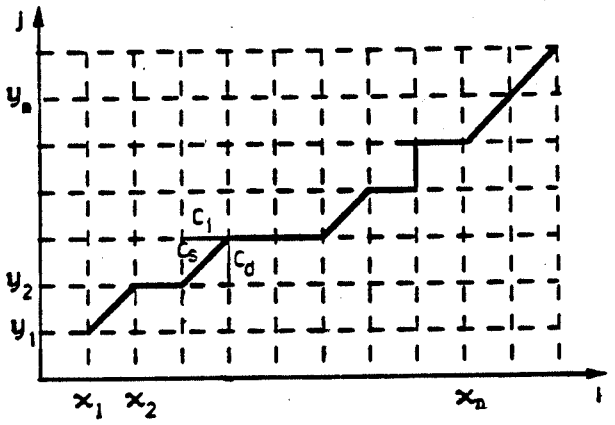


Figure 3: Alignement temporel entre chaîne et modèle phonétique

Cette méthode d'optimisation estime ces fonctions de coût à partir d'un grand nombre de chaînes phonétiques, et de leurs modèles respectifs contenus dans le lexique. Ces chaînes phonétiques sont extraites, lors d'une phase d'apprentissage, à partir de plusieurs locuteurs afin que la matrice de confusion puisse être utilisée pour n'importe quel locuteur.
L'algorithme utilisé est le suivant:

1. Initialiser $p_S(x_i, y_j)$ et $p_D(y_j)$

$$C_S = -\log_{10} p_S(x_i, y_j)$$
$$C_D = -\log_{10} p_D(y_j)$$

2. Initialiser les compteurs :
$N_S(x_i, y_j) = N_D(y_j) = 0$

3. Début des itérations : calculer les fonctions de déviations utilisant les valeurs de p_S et p_D (algorithme de programmation dynamique).

   En faisant un retour arrière sur le chemin optimal, les deux compteurs N_S et N_D sont remis à jour en incrémentant N_S à chaque transition diagonale ou horizontale, et N_D à chaque transition verticale.

   Les nouvelles probabilités estimées sont calculées par les équations suivantes:

$$p_S(x_i, y_j) = \frac{N_S(x_i, y_j)}{(N_S(x_i, y_j) + N_D(y_j))}$$
$$p_D(y_j) = \frac{N_D(y_j)}{(N_S(x_i, y_j) + N_D(y_j))}$$

4. Faire plusieurs itérations à partir de la troisième étape, jusqu'à ce que p_S et p_D ne varient plus significativement, ou fixer un nombre maximum d'itérations. L. Baum [7] a montré mathématiquement que cette mèthode atteint un optimum local.

L'initialisation de p_S et de p_D peut se faire de manière aléatoire. Dans notre cas, nous avons initialiser ces deux matrices par un algorithme analogue au précédent, mais en alignant la chaîne phonétique à son énoncé correspondant segmenté et étiqueté manuellement.
Un lissage des probabilités peut s'avérer utile, surtout s'il n'y a pas suffisament de données. A partir des connaissances à priori sur les phonèmes, nous définissons des classes d'équivalence entre symboles du lexique. Ainsi, la méthode d'interpolation utilisée est la suivante:
Pour chaque classe U_n $(n = 1, \ldots, K)$, et pour tout x_i et $x_l \in U_n$:

$$Q(x_i, y_j) = \sum_{l=1}^{cardU_n} p(x_l, y_j)/cardU_n$$
$$p^*(x_i, y_j) = \lambda_i * p(x_i, y_j) + (1 - \lambda_i) * Q(x_i, y_j)$$

où λ_i est un coefficient d'interpolation, et dépend de chaque phonème.

7 EXPÉRIENCES ET RÉSULTATS

Lors de la phase d'apprentissage, deux listes de mots ont été utilisées: une contenant les 200 mots les plus fréquents, l'autre 300 mots équilibrés phonétiquement. Ces deux listes ont été acquises et segmentées manuellement à partir de 10 locuteurs (5 hommes et 5 femmes). La liste des mots servant pour l'extraction des prototypes est la même pour tous les locuteurs. Ces 500 mots délivrent 5000 chaînes phonétiques pour les 10 locuteurs, et servent à construire une matrice de confusion pluri-locuteur.
Une autre liste de 500 mots contenant les paires de phonèmes les plus fréquentes, et choisie selon des critères de longueur du mot (2 à 3 syllabes), a servi pour tester les performances du module de présélection.
Le lexique contient des modèles de mots, générés à partir du programme GRAPHON de conversion graphème-phonème réalisé au *LIMSI*. Chaque mot peut être représenté par un modèle ou plusieurs selon les régles phonologiques.
Nous avons mené une série d'expériences sur 3 lexiques (1000, 2000 et 8000 mots). Les 1000 mots du premier lexique sont inclus dans les 2000 mots qui eux aussi sont inclus dans les 8000 mots.
Les pourcentages pour que le mot test soit parmi les 40, 100 et 200 candidats délivrés par la présélection sont donnés dans la table 1.
A partir des résultats obtenus, nous pouvons noter que le système n'est pas dépendant du sexe des locuteurs, puisque les résultats sont bien répartis entre les femmes et les hommes. Par contre, nous pouvons noter l'intervalle assez important entre les bons et mauvais résultats. Ceci est dû à la liste des mots utilisés pour extraire les prototypes; comme cette liste est exactement la même pour tous les locuteurs,

	1000		2000		8000		
Hommes	40	100	40	100	40	100	200
HAK	98.6	98.8	98.2	98.8	92.0	96.2	97.6
GIL	97.6	98.8	96.0	98.2	91.6	96.0	97.3
GAL	96.6	98.6	94.6	97.8	84.2	92.6	95.2
FRD	94.6	98.6	79.2	87.0	79.2	87.0	93.4
PHI	91.8	97.0	87.8	94.2	75.2	84.2	90.2
Femmes	40	100	40	100	40	100	200
SAB	98.4	99.4	97.0	98.8	92.8	93.6	97.4
MAD	97.4	99.0	95.6	98.2	90.5	94.8	96.4
MRT	97.2	98.4	95.0	97.2	88.9	92.8	95.8
LOR	93.8	96.8	90.2	95.2	77.2	86.8	92.4
ROK	94.0	96.4	90.3	94.4	75.1	85.4	91.1
Moyenne	96.0	98.2	92.4	96.0	84.7	90.9	94.7

Tableau 1: Résultats de présélection pour 10 locuteurs et trois dictionnaires de 1000, 2000 et 8000 mots

il se trouve que pour quelques locuteurs ces prototypes ne sont pas les meilleurs. Ceci a été confirmé par l'amélioration des résultats des mauvais locuteurs, en prenant comme prototypes la moyenne de plusieurs mêmes prototypes tirés de différents mots. Ces améliorations n'ont pas été reportées dans le tableau des résultats car elles n'ont été faites que pour les quelques mauvais locuteurs, et nous avons tenu à garder les mêmes conditions pour les résultats affichés.
Les 3 lexiques utilisés ne contiennent qu'un modèle par mot. Nous pouvons encore améliorer ces résultats en introduisant les régles phonologiques dans les lexiques.

8 CONCLUSION

Hormis quelques locuteurs ayant relativement de mauvais résultats, nous estimons que les résultats sont en moyenne assez satisfaisants étant données les restrictions déjà mentionnées.
En utilisant des lexiques multi-modèles et en moyennant les prototypes, nous pouvons encore améliorer les performances du module de présélection. Néanmoins, les résultats déjà obtenus fournissent une base suffisament sûre pour les étapes suivantes du système de reconnaissance: le module d'analyse phonétique fine et le module d'analyse contextuelle.

Références

[1] R. Billi, G. Arman, D. Cericola, G. Massia, M. Mollo, F. Tafini, G. Varese, V. Vittorelli , "A PC-Based Large Vocabulary Isolated Word Speech Recognition System.", In *Proceedings* of EUROSPEECH 1989 vol 2, pp 157 – 160.

[2] H. Drexler, R. Roddeman, L. Boves, H. Strik, "Optimizing Lexical Fast Search in a Large Vocabulary Isolated Word Speech Recognition System.", In *Proceedings* of EUROSPEECH 1991 vol 3, pp 1401 – 1404, Genova.

[3] R. Billi, G. Massia, F. Nesti, "Word Preselection for Large Vocabulary Speech Recognition.", Proc. ICASSP, 1986, Tokyo

[4] Le Roux, C. Gueguen, "A fixed point computation of partial correlation coefficients.", IEEE Trans. Acoust., Vol 25, pp 257 – 259, 1976

[5] Shuzo Saito, Kazuo Nakata, "Fundamentals of speech signal processing.", Tokyo, 1985, pp 96 – 97

[6] H. Sugiyama, H. Shikano, "LPC peak weighted spectral matching measures.", Trans. of IECE, Vol $J64 - A, N°5$, 1981

[7] L. Baum, "An Inequality and Associated Maximization Technique in Statistical Estimation for Probabilistic Functions of Markov Processes.", Inéqualities. 1972

19es J.E.P. - Bruxelles, 19 au 22 mai 1992.

RECONNAISSANCE DE VOCABULAIRES DIFFICILES A L'AIDE DE RESEAUX NEURONAUX

Yolande Anglade (1)(2), Dominique Fohr (1) et Jean-Marie Pierrel (1).

(1) **CRIN-CNRS & INRIA Lorraine**
B.P. 239 F54506 Vandoeuvre-lès-Nancy CEDEX
(2) **SOLLAC**
57191 Florange CEDEX

RESUME

Le but de ce travail est d'améliorer la reconnaissance automatique de mots acoustiquement similaires - ne se différenciant que par un seul phonème - à travers l'exemple typique du vocabulaire des lettres de l'alphabet. Notre étude propose une comparaison d'une méthode globale telle que la programmation dynamique et d'une nouvelle méthode basée sur des réseaux neuronaux. Cette méthode a pour principe de rechercher les trames discriminantes qui permettront de distinguer les mots du vocabulaire. Cette recherche peut être basée sur des critères energétiques, spectraux ou temporels selon les mots concernés. Testée dans un contexte monolocuteur, puis multilocuteur, cette méthode nous a permis d'augmenter les scores de reconnaissance de façon très significative dans les deux configurations.

INTRODUCTION

Les performances des méthodes de reconnaissance globales, telles que la programmation dynamique ou les modèles de Markov cachés, donnent maintenant de bons résultats. Cependant, ces méthodes ne permettent pas une bonne discrimination des mots acoustiquement similaires [FAN90] [COL91], mots ne se différenciant que par un phonème. Le vocabulaire des lettres de l'alphabet comporte un certain nombre de telles difficultés. Son utilisation est cependant utile dans de nombreuses applications, en particulier pour l'épellation de noms propres [ANG91]. A la suite d'expériences de reconnaissance, nous avons observé un nombre important d'erreurs pour cinq sous vocabulaires difficiles : (A,K), (P,T), (U,Q), (B,D,V), (L,M,N). Le but de cette étude est d'améliorer les scores de reconnaissance pour chacun de ces sous-vocabulaires. Pour effectuer cette discrimination, nous avons utilisé une méthode basée sur un réseau de neurones. Dans un premier temps, nous allons décrire les caractéristiques de cette nouvelle méthode, puis nous donnerons les résultats en les comparant sur la base de données BDSONS à ceux d'une méthode globale. L'étude a tout d'abord été menée dans un cadre monolocuteur, puis étendue au niveau multilocuteur.

METHODE

1 Discrimination à l'aide de réseaux neuronaux

L'idée de cette nouvelle méthode (voir figure 1) est de sélectionner la partie discriminante des lettres permettant de faire la reconnaissance pour chacun des sous-vocabulaires considérés. Une paramétrisation est ensuite calculée sur cette partie et les vecteurs résultants sont fournis à un réseau neuronal. Nous avons utilisé un réseau neuronal pour chaque sous-vocabulaire En effet, les parties discrimantes sont différentes suivant les phonèmes à distinguer : consonne finale pour (L,M,N), ou initiale pour (P,T) par exemple.

1.1 Analyse phonétique

Les sous-vocabulaires définis se caractérisent par le fait que seul un phonème permette de les distinguer. Il nous faudra donc rechercher la partie caractéristique propre à ce phonème, se situant:

- au niveau de la barre d'explosion et des transitions pour (P,T),
- au niveau des transitions pour (U,Q) et (A,K), et de la présence ou non de la barre d'explosion,
- au niveau de la consonne finale pour (L,M,N),
- à deux niveaux pour (B,D,V):
  - barre d'explosion et transitions pour B et D,
  - consonne initiale pour B et V.

Figure 1 : principe de la méthode

1.2 Recherche de la partie discriminante

Plusieurs méthodes de positionnement des trames utilisées pour la discrimination ont été mises en oeuvre (cf tableau 1). De plus, une ou plusieurs trames peuvent être utilisées pour tenir compte de l'aspect dynamique de certaines caractéristiques phonétiques (par exemple les transitions). Nous allons maintenant détailler ces méthodes :

- méthode énergétique:
  le but de cet procédure, utilisée pour (A,K), (U,Q) et (B,D,V), est de positionner une trame discriminante relativement à un seuil d'énergie. Dans une première phase, on recherche la fenêtre la plus énergétique dans le mot (cette fenêtre sera donc située dans la voyelle). A partir de cette fenêtre, on détermine une fenêtre, située en amont, dont l'énergie est deux fois plus faible. On vient ainsi se positionner à la transition entre la consonne initiale et la voyelle. Le fait d'utiliser un rapport d'énergie pour positionner cette trame, nous permet d'être indépendant du niveau sonore du locuteur.
- méthode spectrale:
  la méthode spectrale, utilisée pour le vocabulaire (P,T), a pour objectif la détermination de la position de la barre d'explosion. Pour ce faire, nous calculons sur chaque fenêtre d'acquisition de 16ms, le maximun M(t) du spectre (obtenu par transformée de Fourrier) dans la bande de fréquence 700–8000Hz. On détermine ensuite un seuil S égal à min(M(t)) + 3dB. L'instant t qui vérifie M(t) > S, est celui qui nous sevira de référence
- méthode temporelle:
  Pour le sous vocabulaire (L,M,N), il nous faut placer la fenêtre d'analyse dans la partie consonantique. Après avoir recherché le début et la fin du mot (par une méthode fondée sur un critère d'énergie), nous avons positionné la trame au $5/8^{ème}$ (valeur expérimentale) du mot.

Sous vocabulaire	Sélection partie discriminante	nombre de trames
(P,T)	spectrale	2
(A,K) (U,Q)	énergie	1
(B,D,V)	énergie	2
(L,M,N)	temporelle	1

Tableau 1 : méthodes utilisées pour rechercher la partie discriminante.

1.3 Paramétrisation

Après sélection de la ou des trames discriminantes, nous calculons des coefficients cepstraux (échelle Mel) sur une fenêtre de 32ms. L'ordre retenu pour cette paramétrisation est de 12; cette valeur a été choisie expérimentalement après une étude concernant son influence sur les performances de reconnaissance (cf résultats, figure 4)

1.4 Caractéristiques des réseaux neuronaux

Ils sont composés d'une couche d'entrée, comprenant une neurone par coefficient cepstral calculé, d'une couche cachée à 6 neurones et d'une couche de sortie avec un neurone par lettre du sous-vocabulaire. Leur apprentissage est réalisé par rétropopagation d'erreur.

2 Méthode globale

Nous avons choisi de comparer les résultats obtenus à l'aide des réseaux neuronaux à ceux obtenus par programmation dynamique. Nous avons utilisé une quantification vectorielle (256 classes), une paramétrisation identique à celle précédemment employée (12 coefficients cepstraux) et une distance euclidienne. Quatre références pour chacune des lettres du vocabulaire ont servi à faire la reconnaissance.

RESULTATS et COMPARAISON

1 Corpus

Le corpus BDSONS du GRECO dans lequel 26 locuteurs (13 hommes, 13 femmes) ont prononcé 16 répetitions de tout l'alphabet en mots isolés, a été utilisé pour comparer les deux méthodes. L'apprentissage a été effectué sur 4 répétitions, le test sur les 12 autres. Pour le test multilocuteur, 7 hommes et 6 femmes ont été selectionnés pour l'apprentissage et les locuteurs restants (6 hommes et 7 femmes) pour les tests.

2 Resultats

2.1 Position de la trame

La recherche de la partie discriminante, décrite ci-dessus, nous a permis de positionner un point de référence dans le mot. Pour étudier l'influence de la position de la trame sur les scores de reconnaissance, nous avons déplacé une fenêtre d'analyse par rapport à ce point de référence. La figure 2 montre les résultats de reconnaissance obtenus pour le vocabulaire (P,T) en multilocuteur. Sur l'axe des abscisses, la position de la trame retenue est indiquée en millisecondes par rapport au point de référence (0). Les traits verticaux situés en chaque point de la courbe correspondent eux à l'intervalle de confiance des scores de reconnaissance. Ces intervalles, calculés statistiquement, dépendent de la valeur du score et du nombre d'échantillons pour lequel il a été obtenu. On observe sur cette figure que la discrimination est maximale autour du point de référence choisi (+/- 8ms). En revanche, si l'on place la trame avant ce point, les performances chutent car on se situe alors dans l'occlusion du P ou du T, qui ne contient aucune information pertinente. De même, si l'on se positionne après ce point, c'est-à-dire dans la voyelle, les performances sont également très mauvaises.

La figure 3 permet également de se rendre compte de l'importance d'une bonne localisation de la trame dans le cadre du vocabulaire (U,Q). Pour discriminer correctement ces deux lettres, il est nécessaire de se placer juste avant le début de la voyelle. En effet, cela correspond dans le cas d'un "U" à la partie silencieuse précédant la voyelle alors que dans le cas d'un "Q" on se situe dans la barre d'explosion. Dans ce sous-vocabulaire, les deux types de trame étant très différentes, la discrimination est très bonne: 99% de bonne reconnaissance.

2.2 Ordre de l'analyse cepstrale

La figure 4 montre l'influence de l'ordre d'analyse sur les scores de reconnaissance. Nous avons fait varier de 8 à 24 le nombre de coefficients cepstraux. Les résultats, obtenus sur le vocabulaire (L,M,N) ne montrent pas une influence significative de ce paramètre sur le taux de reconnaissance. En effet, les variations de cette courbe se situent à l'intérieur de l'intervalle de confiance des résultats.

2.3 Résultats monolocuteur

Le tableau 2 montre les résultats obtenus pour les différents sous-vocabulaires étudiés et chacune des deux méthodes. On constate que les réseaux de neurones apportent une très importante amélioration des scores de reconnaissance pour l'ensemble des sous-vocabulaires, principalement pour (P,T), (B,D,V) et (L,M,N). Ces résultats montrent l'intérêt de considérer uniquement les trames discriminantes de chaque lettre plutôt que de faire une comparaison globale sur l'ensemble de la forme.

Sous-vocabulaires		Program. dynamique	Réseaux neuronaux
P/T	%	80%	97%
	nb erreurs	127/624	16/624
U/Q	%	97%	99%
	nb erreurs	18/624	7/624
A/K	%	93%	99%
	nb erreurs	42/624	1/624
B/D/V	%	75%	88%
	nb erreurs	232/936	112/936
L/M/N	%	62%	93%
	nb erreurs	359/936	65/936

Tableau 2 : résultats obtenus par programmation dynamique et réseaux neuronaux en monolocuteur.

2.4 Résultats multilocuteur

Les résultats obtenus en multilocuteur sont donnés dans le tableau 3. Ils montrent la validité de l'approche étendue dans ce cadre. Nous pouvons constater que les résultats sont à peu près semblables à ceux obtenus en monolocuteur. Dans le cas de (B,D,V), nous pouvons même observer une amélioration des performances, dont l'origine peut être expliquée par un meilleur apprentissage du réseau de neurones.

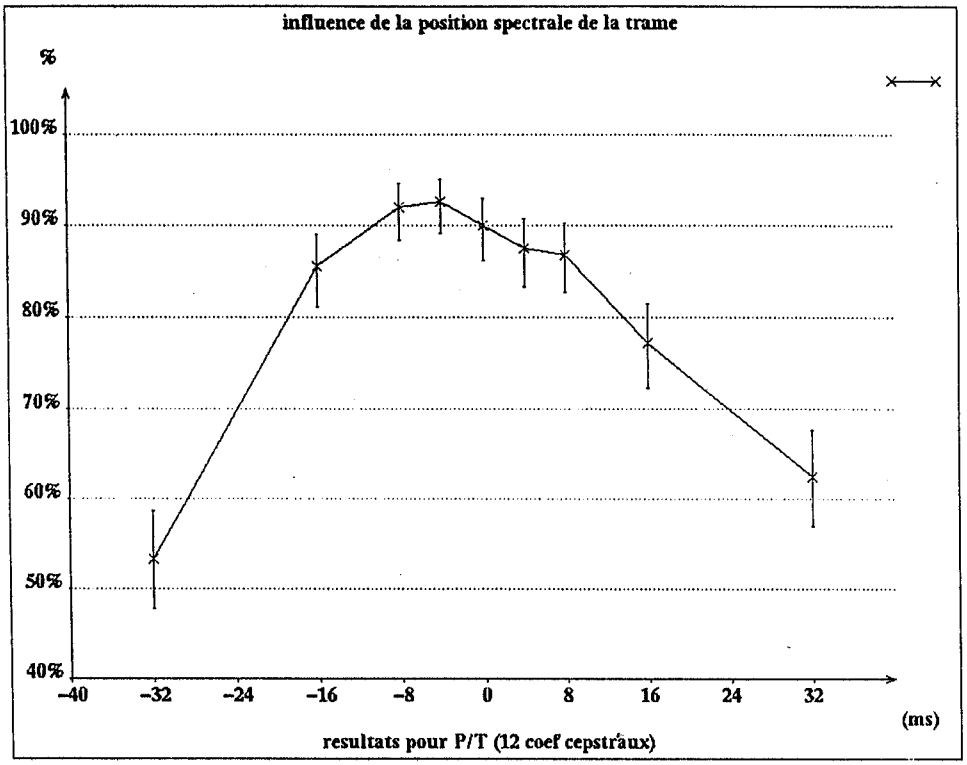


Figure 2 : influence de la position spectrale de la trame pour le vocabulaire (P,T).

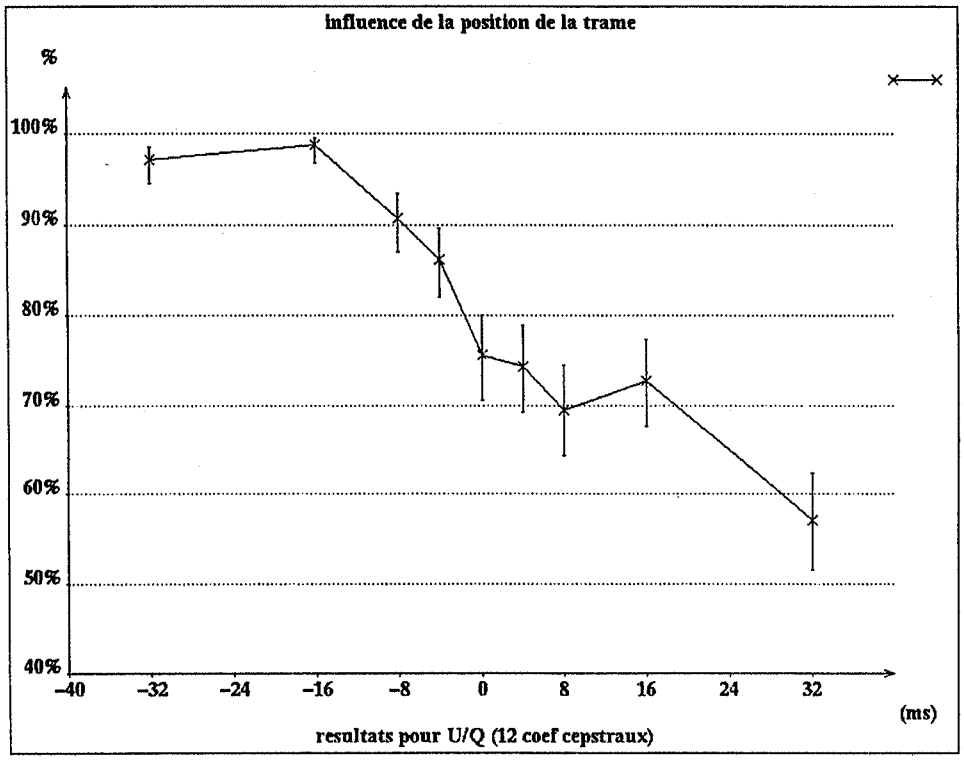


Figure 3 : influence de la position de la trame pour le vocabulaire (U,Q).

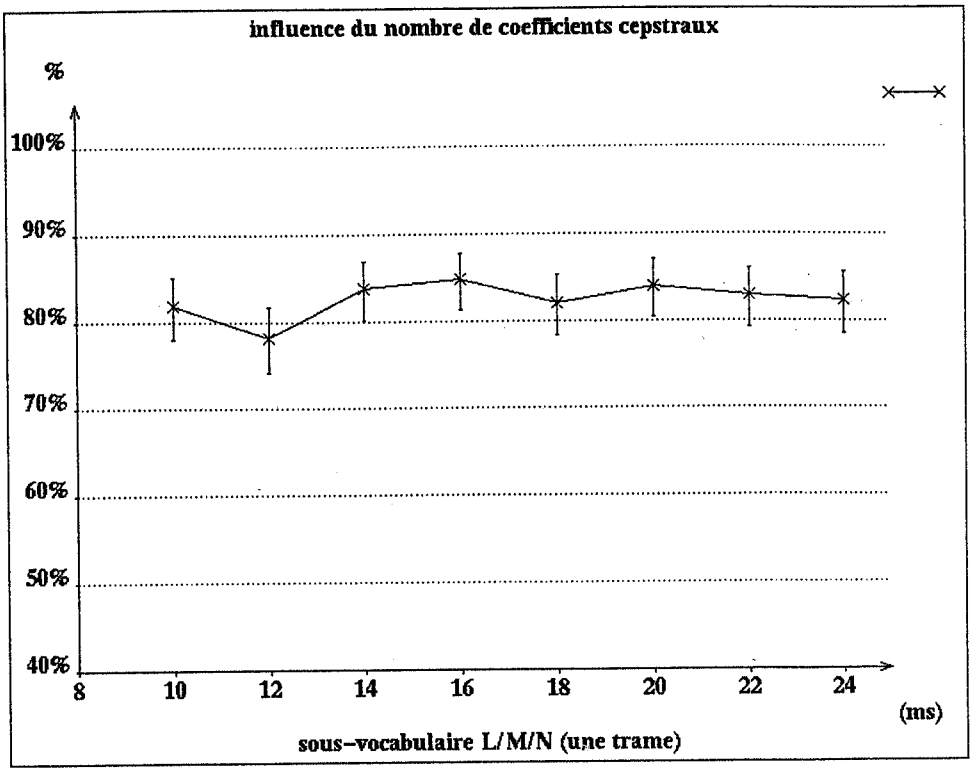


Figure 4 : influence de l'ordre de l'analyse cepstrale pour le vocabulaire (L,M,N)

Sous-vocabulaires		réseaux neuronaux
P/T	%	94%
	nb erreurs	18/312
U/Q	%	99%
	nb erreurs	4/312
A/K	%	99%
	nb erreurs	3/312
B/D/V	%	93%
	nb erreurs	32/468
L/M/N	%	80%
	nb erreurs	96/468

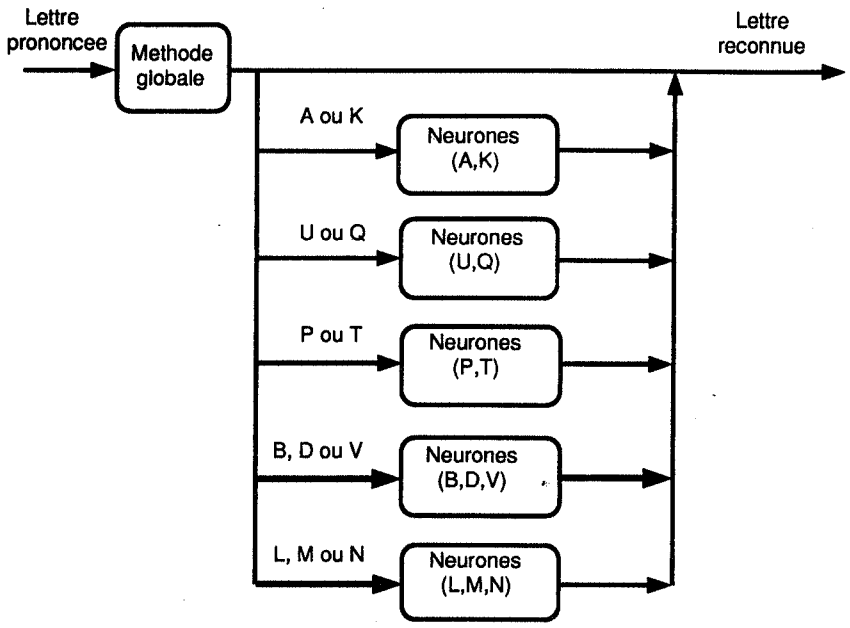
Tableau 3 : résultats obtenus par programmation dynamique et réseaux neuronaux en multilocuteur.

CONCLUSION

La contribution originale de ce travail est l' utilisation des réseaux de neurones pour améliorer les scores de reconnaissance sur des vocabulaires difficiles, composés de mots ne se distinguant que par un seul phonème. Nous avons montré que cette méthode donne de bien meilleurs résultats que la programmation dynamique, tant dans un cadre monolocuteur que multilocuteur. Des algorithmes de localisation des trames discriminatives doivent être définis dans chacun des vocabulaires concernés. Leur précision est un élément important de cette méthode si l'on veut fournir aux réseaux de neurones les informations pertinentes. Notre analyse relative à la position de ces trames montre à quel point cela conditionne les performances de reconnaissance.

Pour intégrer ces améliorations dans un système de reconnaissance dont la tâche consisterait à effectuer une reconnaissance sur les lettres de l'alphabet, une architecture à deux niveaux peut être adoptée (voir figure 5) : un premier étage de reconnaissance globale (programmation dynamique ou modèles de Markov) complété par un deuxième utilisant les réseaux de neurones proposés et déclenchés uniquement en cas de confusions possibles. Un de nos objectifs est maintenant de tester une telle architecture de façon à obtenir des scores de reconnaissance globaux sur ce vocabulaire.

Figure 5 : architecture proposée



Une autre voie de recherche consiste à tester la robustesse de cette méthode dans des environnements bruités. Nous menons actuellement sur ce thème un travail consistant à tester les performances de ces réseaux de neurones, sur de la parole Lombard (prononcée dans du bruit). Les premiers résultats obtenus sont encourageants et semblent indiquer que là-aussi notre méthode apporte une amélioration sensible des résultats de reconnaissance.

BIBLIOGRAPHIE

[ANG91] Y. Anglade, J.M. Pierrel, J.C. Junqua. "A spoken language interface for a telephone switchboard operator center". Proceedings EUROSPEECH 1991, 307–310, Genova Italy 1991.

[COL91] R.A. Cole, M. Fanty, M. Gopalakrishnan, R.D.T. Janssen. "Speaker-independent retrieval from spellings using a database of 50000 names". Proceedings ICASSP 1991, 325–328, Toronto Canada 1991.

[FAN90] M. Fanty, R.A. Cole. "Speaker-independent English alphabet recognition: Experiments with the E-Set." Proceedings ICSLP 1990, 1990.

[GON90] Y. Gong, J.P. Haton. "Towards a General Signal interpretation System signal-to-symbol conversion level. " Proceedings Xth IEEE ICPR 1990, 79–84, Atlantic city USA 1990.

19es J.E.P. - Bruxelles, 19 au 22 mai 1992.

CONTRIBUTION DE RÉSEAUX NEURONAUX POUR LA RECONNAISSANCE DES OCCLUSIVES AU SEIN DU SYSTÈME EXPERT APHODEX

Dominique François et Dominique Fohr

CRIN-CNRS & INRIA Lorraine
B.P. 239 F54506 Vandoeuvre-lès-Nancy CEDEX

Résumé

Notre étude propose une utilisation de réseaux de neurones par un système expert de décodage acoustico-phonétique dans le but d'améliorer la reconnaissance des consones occlusives. Nous présentons le travail effectué jusqu'à présent dans un contexte analytique pour augmenter l'efficacité du système à base de connaissances phonétiques. Et dans un deuxième temps nous proposons une nouvelle méthode reposant sur la conception de perceptrons multi-couches pour l'identification de plosives destinés à être intégrés au système expert. Nous concluerons sur la collaboration des deux methodes dans un système hybride..

INTRODUCTION

Le projet APHODEX a commencé il y a plusieurs années avec l'intention d'apporter une aide au décodage acoustico-phonétique de la parole continue. L'idée était de concevoir un système à base de connaissances s'inspirant de la méthode des phonéticiens pour lire des spectrogrammes. Une première réalisation a pris la forme d'un système expert à règles de production capable d'obtenir de façon entièrement autonome un décodage de phrases parlées à partir d'un signal numérisé. Ce système donne de bons résultats [FRA90], toutefois certaines erreurs de reconnaissances peuvent encore être évitées. En particulier en ce qui concerne les occlusives dont l'analyse acoustique est des plus difficiles. Les ereurs plus facilement détectables sont celles dues à une mauvaise détection des barres d'explosion.

MÉTHODE ANALYTIQUE

1 Détection de la barre d'explosion

1.1 Présentation

Pârmi les erreurs constatées au sein de l'identificaton des consonnes occlusives, les erreurs les plus faciles à cerner sont celles survenant lors de la detection des barres d'explosion dans la phrase prononcée. En effet au cours d'un décodage, les affichages en surimpression sur le spectrogramme ne permettent pas le contrôle de toute l'analyse acoustique. Par contre, la bonne localisation d'une barre d'explosion présente sur le spectrogramme peut être controlée de facon assez précise. Il paraissait clair que APHODEX avait besoin d'une détection de burst plus robuste. On comprendra d'autant mieux ce besoin que le burst est l'évenement acoustique le plus caractéristique des occlusives. Certains indices étant recherchés à l'intérieur d'une période extrèmement breve, l'attaque du burst, il va de soi que la localisation de celle-ci doit être des plus précise.

1.2 Principe

Le principe repose sur le calcul d'une distance entre une forme moyenne de barre d'explosion et une zone du spectrogramme à un instant donné. En réalité , huit distances sont évaluées, une par bande de fréquence de 1000 Hz. Pour chaque bande de fréquences, on calcule sur une zone de 392 millisecondes de large une fenêtre de 7 valeurs d'énergie, ceci à partir du spectrogramme global de la phrase obtenu par une FFT d'ordre 256. La forme moyenne moyenne d'une barre d'explosion est décomposée de la même façon sur 8 bandes de fréquences et en 7 valeurs temporellement. A chaque instant du spectrogramme, pour chaque bande de fréquences, la distance est égale à la différence cumulée de la fenêtre de référence et de la fenêtre d'énergies calculées.

On obtient ainsi 8 courbes de distances qui présentent des creux plus ou moins importants aux instants correspondant aux barres d'explosions présentes sur le spectrogramme. La figure suivante montre un spectrogramme et la courbe moyenne des 8 distances calculées. Un algorithme de détection de pics dont les seuils ont été ajustés expérimentalement détermine les zones du spectrogramme succeptibles de contenir un burst. La majorité des plosives dénotant une absence nette d'explosion sont à ce moment déja rejetées. Les autres font alors l'objet d'une analyse plus fine. Pour ce faire on calcule sur la zone précedement déterminée un petit spectrogramme par une transformée de Fourrier d'ordre 64. Les variations d'énergie qui suivent immédiatement le silence de la plosive nous permettent alors de distinguer une barre d'explosion d'une forte attaque en début du contexte droit. La régularité et la continuité dans l'intensité des pics présents infirmeront la présence d'une barre d'explosion, ainsi qu'une dérivée spectrale faible de pic à pic.

2 Analyse

L'analyse du système expert débute par une phase de segmentation du spectrogramme en une suite discrète de phonèmes à identifier. Cette procédure fait en même temps une classification en 4 categories: voyelles, occlusives, fricatives et sonnantes. Un ensemble de règles est dédié à chaque catégorie. Des règles ont été ecrites à partir des nouvelles procédures d'extraction d'indices. La présence et l'intensité de la barre d'explosion sont issues de la procédure décrite précédement. D'autres indices sont extraits du burts comme les fréquences des maxima d'énergie, la forme génerale du spectre (ascendante, descendante, compacité..) et des rapports d'energie de certaines bandes de fréquences déterminées experimentalement. Chaque règle contibue à la construction d'un treillis de phonèmes pondérés par des coéfficients de plausibilité. Nous présentons dans le dernier chapitre l'évaluation de treillis obtenus à partir d'un corpus de parole continue manuellement étiquetée.

MÉTHODE CONNEXIONISTE

1 Paramétrisation

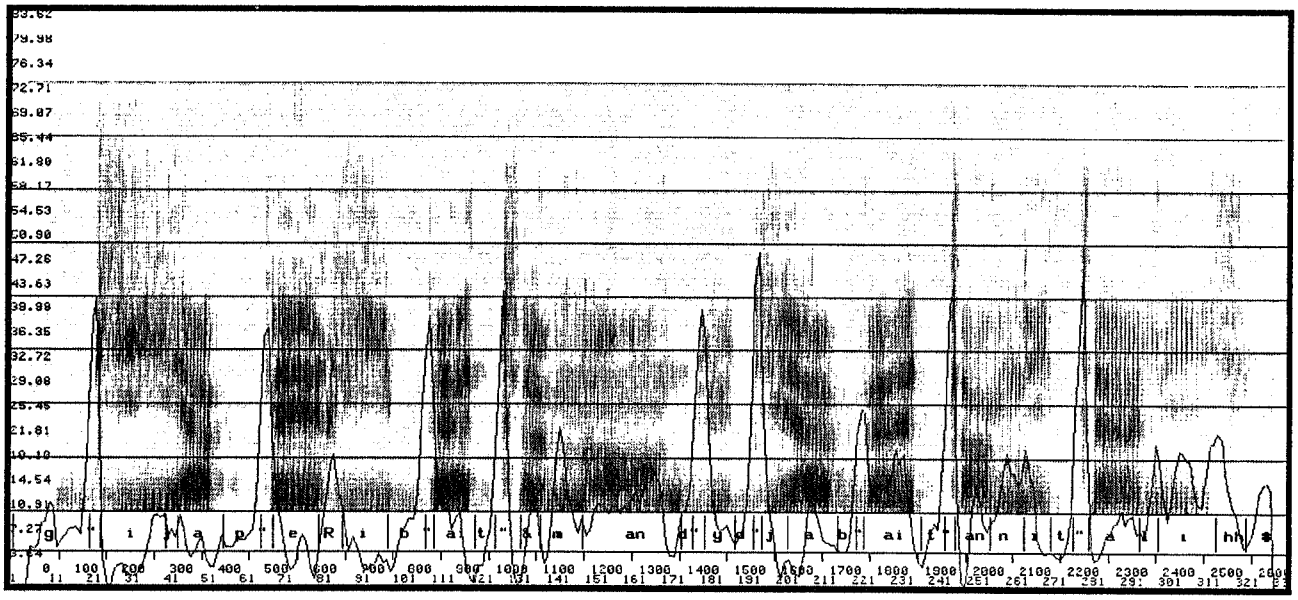
L'idée de cette méthode est d'utiliser des réseaux de neurones afin de résoudre les difficultés rencontrées lors de la phase de modèlisation de l'expertise. Les réseaux utilisés acceptent en entrée des vecteurs représentant le spectre d'une barre d'explosion. Ainsi la méthode connexioniste et le système expert, dans son analyse du burst, traitent des données de même nature. Après localisation de la barre d'explosion dans la plosive présentée, des coefficients cepstraux en echelle Mel sont calculés à partir d'une fenêtre de 32 millisecondes.

2 Architecture

L'architecture mise en œuvre est celle des perceptrons multicouches. Chaque perceptron est composé de trois couches:

- Une couche d'entrée comportant 48 cellules, une par coefficient cepstral.
- Une couche cachée de 6 neurones.
- Une couche de sortie comprenant une cellule par phonème à identifier.

Le nombre de cellules en entrée a fait l'objet d'une étude expérimentale, une représentation cepstrale d'ordre 48 a donné les meilleurs résultats. L'apprentisage des perceptrons s'est fait par la technique de rétropropagation d'erreurs. On pourra consulter [GON90].



RESULTATS EXPÉRIMENTAUX

1 Corpus

Pour nos expérimentations nous avons utilisé deux corpora de parole continue:

- Le corpus BDSONS du GRECO "La bise et le soleil", il s'agit d'un texte lu par 8 locuteurs.
- Un corpus de parole continue contenant 4 répétitions de 17 phrases prononcées par 18 locuteurs masculins, ce qui représente 40330 phonèmes.

En ce qui concerne les réseaux de neurones, les apprentissages ont été effectué à partir de 9 locuteurs et les tests à partir de 9 autres.

2 Resultats

2.1 Détection de la barre d'explosion

L'algorithme de détection de barres d'explosion a un taux de bonne localisation de 95 %. Les taux d'erreur sont : insertion 3,58 %, omission 0,06 % et confusion 1,36 %

2.2 Système à base de connaissances.

L'evaluation des treillis obtenus par le système expert seul est présentée par des matrices de confusions. La matrice montrant les performances d'identification des plosives est présentée en Figure 1.

2.3 Réseaux de neurones

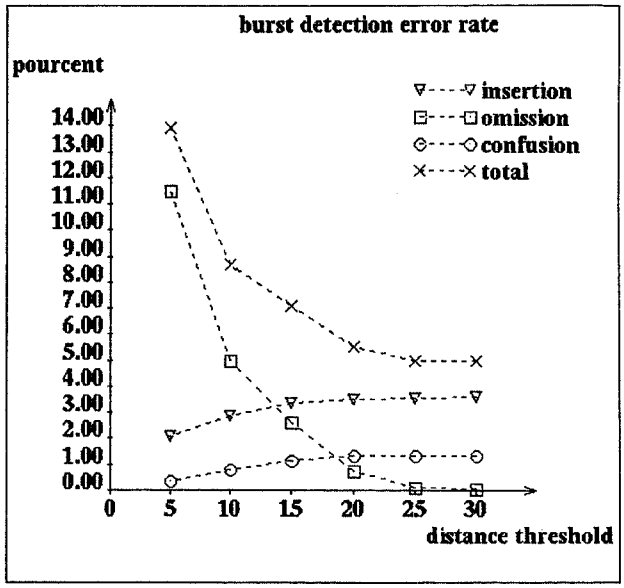
Les résultats présentés ici ont été obtenus avec les contextes les plus nombreux existant dans le corpus. Les occlusives ont été extraites de celui-ci à l'aide de l'étiquetage manuel, la barre d'explosion étant localisée par la même procédure que celle décrite précédement. Le Tableau 1 présente les résultats obtenus par les perceptrons mis en œuvre.

Les taux d'identification sont tous très satisfaisants, ils sont supérieurs à 90 % excepté pour pour la discrimination des plosives sourdes en contexte vocalique /i/. Ceci est probablement du au fait que /i/ est une voyelle d'arrière, ceci rendant plus difficile la bonne reconnaissance par l'analyse de la seule barre d'explosion. Une amélioration pourrait être la prise en compte d'une période plus longue pour calculer le vecteur d'entrée du perceptron. Dans ce cas le système analytique offre la possibilité d'étudier le bruit suivant l'explosion ou les transitions si le bruit n'est pas trop important.

CONCLUSION

L'idée originale de ce travail réside sur le fait qu'APHODEX intégre maintenant une méthode qui n'est pas de la famille des méthodes analytiques. Cette méthode à base de réseaux neuronaux a l'avantage de ne pas necessiter une formalisation de connaissances, le perceptron "apprend" de façon autonome. De cette manière on peut résoudre les problèmes relatifs à l'acquisition de l'expertise. En effet, le principal inconvénient d'une méthode analytique est la difficile et laborieuse étape de mise au point qui doit se faire en collaboration avec un expert phonéticien. Par contre, un système expert a l'avantage d'être évolutif et d'admettre des corrections sur sa manière de fonctionner,en particulier l'extraction d'indices, les règles utilisées et la stratégie mise en œuvre. La seule façon d'agir sur le perceptron est la paramétrisation et la détermination d'un corpus d'apprentissage. Ce dernier doit avoir une taille importante, et c'est là qu'on est confronté à une contrainte majeure, le manque de corpus français de grande taille en parole continue.

Malgré cela, les taux de reconnaissance sont très encourageants et sont suffisament éleves pour permettre une amélioration globale des performances de notre système expert. Nos travaux actuels consistent à intégrer les réseaux neuronaux au sein d'APHODEX. Des règles contextuelles permettent de déclancher une identification par un perceptron, cet appel à un réseau de neurones doit se faire selon une stratégie adéquate, soit pour confirmer et renforcer une hypothèse, soit palier une indécision ou encore combler un manque d'indices caractéristiques. Les conflits que peuvent apporter une seconde méthode doivent être résolus au mieux, pour ce faire nous travaillons à l'introduction de méta-règles. C'est de la bonne cohabitation des méthodes que va dépendre le succès d'un tel système hybride.



	Number	p	t	k	b	d	g	Deletion	%
p	498	377	29	0	6	58	13	12	75.7
t	889	55	503	34	201	10	6	66	56.6
k	560	34	49	308	89	12	0	55	55.0
b	471	4	16	6	367	10	2	26	77.9
d	988	93	12	1	156	459	7	153	46.5
g	387	33	14	0	42	15	151	81	39.0
Insertion		40	9	0	21	5	0		

Figure 1 Matrice de confusion évaluant l'identification des occlusives par le système expert.

contexte	*plosives*	*nombre d'occurences*	*taux de reconnaissance*
/o/ / /u/	/p/, /t/, /k/	144	95 %
	/b/, /d/, /g/	95	92 %
/l/	/p/, /k/	44	90 %
/r/	/p/, /t/	28	90 %
/i/	/p/, /t/, /k/	74	81 %
	/b/, /d/	44	97 %
/e/ / /	/p/, /t/	82	90 %
	/b/, /d/	65	95 %
/ /	/p/, /t/, /k/	55	93 %
/a/	/b/, /d/, /g/	105	90 %

Tableau 1 Résultats obtenus par les différents réseaux de neurones.

BIBLIOGRAPHIE

[FRA90] D. Françcois, D. Fohr. "Première évaluation d'APHODEX, système expert pour le décodage acoustico-phonétique de la parole continue." 19émes J.E.P., Montréal CA 1990.

[GON90] Y. Gong, J.P. Haton. "Towards a General Signal interpretation System signal-to-symbol conversion level. " Proceedings Xth IEEE ICPR 1990, 79–84, Atlantic city USA 1990.

19es J.E.P. - Bruxelles, 19 au 22 mai 1992.

Utilisation des méthodes de raisonnement hypothétique en reconnaissance de la parole continue

Anne BONNEAU, François CHARPILLET, Sylvie COSTE, Jean-Paul HATON, Yves LAPRIE, Pierre MARQUIS

CRIN-CNRS et INRIA-Lorraine
B.P. 239, 54506 Vandœuvre-lès-Nancy Cedex, FRANCE

Résumé

Malgré des progrès importants effectués au cours de la dernière décennie, le problème de la reconnaissance automatique de la parole n'est pas encore totalement résolu, notamment dans le cas de phrases prononcées de façon continue et spontanée.
Le but de cet article est de montrer comment une telle démarche analytique de reconnaissance peut tirer profit des méthodes de raisonnement hypothétique issues de l'étude des systèmes à bases de connaissances. Pour illustrer ce point, nous nous limiterons ici à l'étape de décodage acoustico-phonétique d'un système général de reconnaissance des phrases.
Nous présentons enfin les résultats expérimentaux obtenus avec plusieurs locuteurs masculins et féminins.

1. Introduction

Malgré des progrès importants effectués au cours de la dernière décennie, le problème de la reconnaissance automatique de la parole n'est pas encore totalement résolu, notamment dans le cas de phrases prononcées de façon continue et spontanée. En fait, les spécificités du signal de parole (variabilité, continuité, phénomènes de co-articulation, etc.) font de la reconnaissance de la parole un des problèmes les plus difficiles auxquels s'attaque l'intelligence artificielle.
Une approche possible consiste à considérer la compréhension d'une phrase inconnue comme un processus s'appuyant de façon implicite sur un ensemble de connaissances organisées hiérarchiquement depuis le niveau acoustique jusqu'au niveau pragmatique, de façon à réduire l'indéterminisme présent aux différents niveaux du traitement. C'est cette approche qui a été adoptée dans le présent travail. Le but de cet article est de montrer comment une telle démarche analytique de reconnaissance peut tirer profit des méthodes de raisonnement hypothétique issues de l'étude des systèmes à bases de connaissances. Pour illustrer ce point, nous nous limiterons ici à l'étape de décodage acoustico-phonétique d'un système général de reconnaissance des phrases.
Dans notre système, les connaissances sont représentées par des modèles phonétiques décrits à l'aide de traits acoustiques par un phonéticien. Lorsque ces traits sont suffisamment forts -à la fois en qualité et en signification- des modèles simplifiés suffisent. Le processus d'identification ressortit à un schéma abductif/déductif (Marquis, 1991) consistant à trouver le ou les modèles qui expliquent au mieux les traits issus du signal acoustique. Comme ce processus peut produire plusieurs réponses, éventuellement incohérentes, un mécanisme global de maintien de cohérence est de plus nécessaire. Les systèmes de type ATMS (De Kleer, 1986) fournissent l'ensemble des fonctionnalités précédentes. C'est la raison pour laquelle nous avons implanté notre système de reconnaissance à l'aide d'un outil logiciel incluant un ATMS, en l'occurrence X-TRA, développé dans notre équipe (Charpillet et al., 1991a).
Il existe relativement peu de travaux analogues dans la littérature . On peut citer Fox et Josephson (Fox, 1991) qui ont proposé une architecture multi-niveaux pour la reconnaissance de la parole fondée sur le modèle de raisonnement abductif par couches. Ils ont pour l'instant implanté le système CV qui reconnait des paires (occlusive-voyelle) par un raisonnement abductif fondé sur des traits acoustiques.
De Mori et Mong ont étudié l'utilisation de plans et de techniques TMS en décodage acoustico-phonétique (De Mori et Mong, 1984), mais dans une optique nettement différente. Deux autres projets japonais utilisent des mécanismes de maintien de vérité et de cohérence dans les

bases de connaissances. Le premier (Komatsu, 1991) a implanté un TMS pour assurer la cohérence des interprétations partielles d'une phrase produites par un ensemble de sources de connaissances dans une architecture de type tableau noir (blackboard). Des essais ont été menés avec une première version comprenant trois sources de connaissances (prosodie, phonétique et "langage"). Le second projet (Nishioka, 1991) se propose d'utiliser un système de résolution de problèmes avec stratégie de recherche de solutions pour implanter un système de compréhension de phrases. Un ATMS a été utilisé par les auteurs pour l'implantation pratique de la stratégie de conduite de l'interprétation.
L'organisation de cet article est la suivante. Après un rapide exposé du problème, nous décrivons les traits acoustiques utilisés pour l'identification des consonnes occlusives en contexte vocalique et les méthodes pour extraire ceux-ci du signal. Nous présentons ensuite notre modèle de reconnaissance par raisonnement hypothétique et nous commentons enfin les résultats expérimentaux obtenus avec quelques locuteurs.

2 Définition du problème

Nous nous limiterons à l'utilisation du raisonnement hypothétique pour la reconnaissance des occlusives. Notre approche, analytique, repose sur la détection d'indices acoustiques pertinents à partir des spectrogrammes de parole.
Nous présentons ci-dessous les principaux indices des occlusives et la stratégie proposée pour leur identification. Un exemple illustre notre méthode.

2.1 Identification des occlusives

Le français possède trois lieux d'articulation pour les occlusives: labial (/p,b/), dental (/t,d/) et palato-vélaire (/k,g/).
Les principaux indices de ces consonnes sont:
- les trajectoires formantiques, plus particulièrement au début ou à la fin de voyelle adjacente,
- les caractéristiques du bruit -composé de la barre d'explosion et du bruit de friction-.

Pour une liste détaillée se reporter à (Edwards, 81).
Afin de prendre en compte les phénomènes de coarticulation, la stratégie d'utilisation des indices ainsi que les indices détectés à droite de la consonne sont définis en fonction de la classe de la voyelle qui suit la consonne. Les transitions formantiques situées à gauche de la consonne sont interprétées en fonction de la classe de la voyelle précédente.
Nous avons distingué trois classes de voyelles : les voyelles antérieures, les voyelles centrales et les voyelles d'arrière.
Lors de l'identification des sons, nous sommes confrontés notamment à trois problèmes :
- l'incertitude des données qui résulte du manque de fiabilité des détecteurs d'événements acoustiques (suivi des transitions formantiques, détecteur de bruits d'explosion),
- l'absence éventuelle de certains indices acoustiques,
- le fait que le pouvoir de discrimination d'un indice dépend toujours de la production du son et n'est donc pas prévisible.

Si, lors de leur réalisation, les indices sont bien détectés et possèdent un pouvoir de discrimination suffisamment élevé, ils sont utilisés comme indices de "préférence" ou d'"exclusion". Un indice de "préférence" permet l'identification d'un son ou d'un trait d'articulation, un indice d'"exclusion" permet l'élimination de certains candidats à l'identification. Le but de tels indices est l'identification (ou le rejet) des sons dans la mesure où ils ne se contredisent pas mutuellement. En cas d'absence d'indice de "préférence", le raisonnement repose sur des indices moins fiables ou moins discriminants. En cas d'émergence de contradiction entre des indices de "préférence" et/ou des indices d'"exclusion", le principe de notre méthode est de maintenir la cohérence des hypothèses ; nous cherchons alors à remettre en question le contexte dans lequel les hypothèses ont été formulées (segmentation des sons, sorties des détecteurs acoustiques, portée du phénomène de coarticulation). Ainsi dans l'exemple du paragraphe 4.2 aucune solution cohérente ne se dégage du fait de la présence simultanée de deux indices forts contradictoires. La remise en cause du contexte peut permettre de trouver la bonne solution : une suite de deux occlusives.

2.2 Détection d'indices acoustiques

Notre approche suppose que l'extraction des indices acoustiques satisfait les conditions suivantes. Un détecteur d'indice fort (un indice de préférence ou un indice d'exclusion) ne doit pas donner lieu à de fausses alarmes. Il doit donc être très robuste, peu sensible et son implantation doit reposer sur des caractéristiques assez grossières du bruit d'explosion. En contrepartie, s'il existe un indice faible correspondant à un indice de préférence son détecteur doit être plus sensible pour couvrir les cas où l'indice acoustique existe mais n'est pas assez prononcé pour que le détecteur de l'indice fort le trouve.

Les détecteurs d'événements acoustiques ont été implantés dans Snorri (Fohr, 1989). Nous ne décrirons ici que les détecteurs destinés à reconnaître qu'une barre d'explosion est celle d'un /k/ suivi d'une voyelle d'arrière. Les caractéristiques du bruit d'explosion les plus significatives quant au lieu d'articulation de l'occlusive correspondent en général à la barre d'explosion. Il faut donc décomposer le bruit d'explosion en deux parties : la barre d'explosion et le bruit de friction.

2.2.1 Décomposition du bruit d'explosion

En général on distingue assez clairement sur les spectrogrammes des bruits d'explosion deux parties :
- la barre d'explosion, éventuellement scindée en deux dans le cas d'une barre d'explosion double,
- le bruit de friction dont les caractéristiques ne nous servent pas.

Pour pouvoir séparer finement la barre d'explosion du bruit de friction nous avons choisi de calculer le

spectrogramme avec une fenêtre temporelle de 4 ms et un déplacement entre deux spectres de 1 ms. Notre décomposition repose sur la modélisation suivante (Fig. 1).

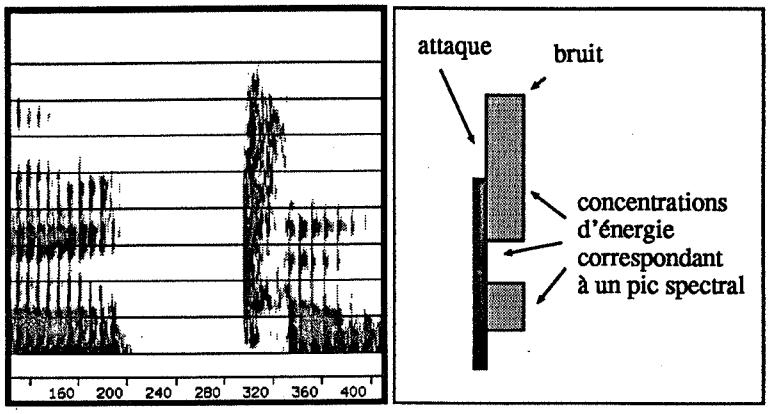


Figure 1 : spectrogramme et décomposition d'un bruit d'explosion.

La barre d'explosion (resp. le bruit de friction) est représentée par les concentrations d'énergie correspondant aux pics spectraux du spectre moyen de la barre d'explosion (resp. le bruit de friction). Seuls les pics suffisamment intenses sont retenus.
La détermination de la frontière se fait en recherchant le point où les modèles de la barre d'explosion et du bruit de friction maximisent un critère de ressemblance avec le bruit d'explosion à analyser. Nous avons testé plusieurs types de critères : énergie expliquée par les pics spectraux des modèles, énergie inexpliquée par les pics du modèle, corrélation des spectres de la barre d'explosion avec le modèle, "remplissage" des modèles, c'est-à-dire l'énergie représentée par le modèle rapportée à l'énergie effective du spectrogramme. C'est le critère de remplissage qui présente le meilleur compromis entre sensibilité et erreur de localisation.

2.2.2 Indices vélaires (dans le contexte /k/ suivi d'une voyelle d'arrière)

Il existe deux indices permettant de déterminer le lieu d'articulation vélaire :
- un indice de préférence appelé "belle barre d'explosion vélaire",
- un indice plus faible appelé "barre d'explosion vélaire".

Ces indices sont évalués à partir de la modélisation de la barre d'explosion.
L'indice "belle barre d'explosion vélaire" est défini par la conjonction des événements acoustiques suivants :
- le pic spectral le plus intense est entre 750 et 1.500 Hz, et il représente plus de 50% de l'énergie située au-dessus de la moyenne du spectre,
- le pic spectral doit apparaître à une fréquence proche de celle de F2 (évaluée par LPC au début de la voyelle suivant l'occlusive),
- la diffusion de ce pic doit être suffisamment faible.

Nous avons défini la diffusion d'un pic spectral comme son moment d'inertie par rapport à la fréquence du maximum spectral rapportée à l'énergie de ce pic.
Un indice de préférence, "belle barre d'explosion vélaire" par exemple, correspond à la présence manifeste sur le spectrogramme d'événements acoustiques caractéristiques d'un lieu d'articulation donné. Nous avons donc imposé que les détecteurs d'indices de préférence retournent une valeur dans l'ensemble {Vrai, Faux}. Ce choix est d'autant plus naturel que ces indices permettent d'orienter fortement la stratégie de décodage ce qui serait délicat si ces indices retournaient une plausibilité choisie dans l'intervalle [0, 1].
L'indice "barre d'explosion vélaire" est défini par la conjonction des événements acoustiques suivants :
- un pic spectral entre 750 et 1.600 Hz,
- ce pic domine la partie du spectre entre 2.000 et 3.500 Hz,
- l'énergie de ce pic est plus importante que celle des autres pics du spectre.

A la différence de l'indice "belle barre d'explosion vélaire" la détection de ces événements comme celle de l'indice lui-même retourne une plausibilité choisie dans [0, 1].
Nous avons vérifié la validité de tous les indices portant sur la barre d'explosion sur 90 occlusives sourdes suivies d'une voyelle d'arrière et extraites du corpus de parole continue «La bise et le soleil ...» de BDSONS. Les indices de préférence se déclenchent dans 86% des cas et il n'y a aucune fausse alarme. Dans 11% des cas l'indice faible permet de préférer la bonne occlusive. Il y a seulement 3% des cas où la bonne occlusive n'est pas classée en tête.

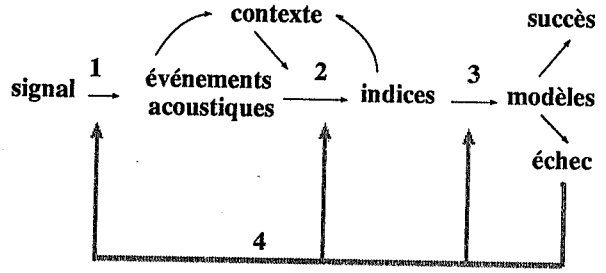
3 Vers une nouvelle méthode de décodage acoustico-phonétique

Dans cette partie, nous décrivons l'utilisation d'un ATMS pour le décodage acoustico-phonétique. Nous présentons tout d'abord le fonctionnement du système réalisé pour le décodage des occlusives, puis nous exposons le schéma de raisonnement abductif-déductif mis en œuvre pour trouver la meilleure explication aux indices acoustiques extraits du signal.

3.1 Description du système

Le système reçoit en entrée les données fournies par les détecteurs acoustiques. Ces données sont les événements acoustiques de l'occlusive et des voyelles adjacentes pour prendre en compte les phénomènes de coarticulation. Les événements acoustiques utilisés sont ceux décrits au paragraphe 2.2.
La première étape consiste à extraire les indices acoustiques à partir des événements et du contexte. Les premières hypothèses émises concernent donc le type des voyelles du contexte. La classe de la voyelle suivante détermine quels indices vont être pertinents et donc indique la stratégie de recherche des indices. Dans tous les cas, les indices négatifs sont pris en compte et une hypothèse est créée correspondant à chaque indice négatif. D'autres hypothèses sont émises correspondant aux indices de préférence trouvés en accord avec le contexte droit. Si aucun indice de préférence n'a été

trouvé, tous les indices sont recherchés et une hypothèse-indice est créée. Dans le cas où un indice de préférence est trouvé, les autres indices ne sont pas recherchés.
La seconde étape consiste à expliquer les indices trouvés précédemment. Nous avons défini pour chaque lieu d'articulation un modèle qui dépend du contexte. Ce modèle est constitué par un ensemble d'indices acoustiques, et est incohérent avec d'autres indices (indices négatifs pour ce lieu d'articulation). Un indice négatif est contradictoire avec un lieu d'articulation dans un contexte donné : par exemple, aucune concentration d'énergie à la hauteur de F2 est contradictoire avec une palato-vélaire suivie d'une voyelle d'arrière. Les problèmes exposés en partie 2.1 montrent que le décodage acoustico-phonétique ne peut pas être abordé par une approche purement déductive. Nous devons trouver une explication aux indices observés. Cependant, une correspondance complète entre les indices du modèle et les indices observés n'est pas envisageable, et nous obtenons en général plusieurs modèles partiellement expliqués. De manière à supprimer certains de ces modèles, nous utilisons ensuite une étape déductive : un indice négatif supprime le modèle avec lequel il est contradictoire. Le processus de décodage suit donc un schéma de raisonnement abductif-déductif.
Le processus de décodage a réussi s'il peut proposer au moins une solution. Il faut remarquer que cette solution peut être mauvaise sans être incohérente. Une telle erreur ne peut pas être détectée par le module de décodage mais seulement par un des niveaux supérieurs (en particulier le niveau lexical).



1. Détecteurs acoustiques : segmentation, suivi de formants, analyseur de la barre d'explosion ...
2. Extraction des indices avec utilisation de critères de préférence
3. Identification des occlusives
4. Retour à une étape précédente du décodage

Figure 2 : principe du décodage.
Le système peut ne trouver aucune solution : soit aucun indice n'a été trouvé, soit tous les modèles proposés après la phase déductive ont été supprimés par des indices négatifs. Nous avons dégagé plusieurs causes :
(i) Une erreur des détecteurs acoustiques a conduit
(a) à supposer artificiellement un indice négatif, le bon modèle a alors été supprimé
(b) à l'absence totale d'indices acoustiques.
(ii) La réalisation de l'occlusive n'est pas conforme aux modèles proposés. Plusieurs possibilités sont à prendre en compte : le son n'est pas une occlusive ou est une suite de deux occlusives ... D'autre méthodes doivent être utilisées pour trouver une solution, ces méthodes n'ont pas encore été définies.
Nous allons travailler sur la définition de ces méthodes. Nous avons déjà réalisé un module correcteur de formants qui suppose de nouvelles valeurs pour les fréquences et les pentes des formants quand la situation (i) est rencontrée. Ces valeurs sont calculées à partir des extrema des spectres et des racines de la LPC.
La figure 2 résume l'ensemble de la méthode.

3.2 Une approche abductive / déductive pour l'interprétation de données symboliques

De façon générale, interpréter des données peut être considéré comme un processus consistant à associer des objets primitifs (ou données) à leurs classes respectives (également appelés concepts, interprétations ou modèles). Par exemple, diagnostiquer l'état du système à partir de ses symptômes relève de l'interprétation de données.

Nous supposons ici que la connaissance du domaine disponible est de type logique. Plus précisément, nous ne considérons que des relations de forme implicative entre données et modèles, représentées dans le cadre de la logique propositionnelle.
Nous nous appuyons sur des méthodes symboliques. La plus simple d'entre elles est totalement déductive ; elle consiste à associer un modèle à une donnée lorsque la représentation de celle-ci constitue une conséquence logique de celui-là. Malheureusement, le raisonnement déductif, pur et simple, est souvent insuffisant pour supporter l'interprétation de données car il est (en général) impossible de décrire complètement, en terme de conditions nécessaires et suffisantes, les concepts du "monde réel", comme les oiseaux, les chaises ... ou les sons.
Cependant, lorsqu'on dispose de conditions nécessaires (mais pas suffisantes) ou suffisantes (et non nécessaires) pour décrire de tels concepts, il peut être intéressant de déterminer la validité de ces conditions car :
- une condition suffisante d'appartenance à une classe C, lorsqu'elle est satisfaite, permet d'établir l'appartenance à C,
- une condition nécessaire d'appartenance à une classe C, lorsqu'elle est satisfaite, constitue une évidence à partir de laquelle on peut supposer l'appartenance à C.

Notre approche consiste à interpréter les données symboliques en nous appuyant sur ces deux familles de conditions. Deux schémas de raisonnement sont conjointement utilisés : l'abduction et la déduction. L'abduction suggère des modèles expliquant tout ou partie des données observées et la déduction permet de supprimer les modèles incompatibles avec l'ensemble des données observées.
La suite de ce paragraphe constitue une description brève et informelle de notre approche, sur un exemple d'école. A titre de pré-requis, nous supposerons que le lecteur a connaissance des principes des raisonnements abductif et déductif (de plus amples détails et quelques points techniques figurent dans (El Ayeb *et al.*, 1990),

(Marquis, 1991)). Soit la connaissance suivante, liant les modèles M1, M2 et M3 aux données d1, d2, d3 et d4:

"si M1 est présent alors d1 et d2 sont présents ;
si M2 est présent alors d1, d3 sont présents mais d2 ne l'est pas ;
si M3 est présent alors d1 et d4 sont présents".

Cette connaissance peut être représentée par la théorie propositionnelle suivante :

$\{(M1 \Rightarrow (d1 \wedge d2)),$
$(M2 \Rightarrow (d1 \wedge \neg d2 \wedge d3)),$
$(M3 \Rightarrow (d1 \wedge d4))\}$

où Mi (i = 1...3) représente l'énoncé atomique "Mi est présent" et di (i = 1...4) représente l'énoncé atomique "di est présent".

Supposons maintenant que les données observées sont d1, d2 et d3. Nous pouvons alors affirmer que :

- les trois modèles M1, M2 et M3 permettent d'expliquer abductivement d1,
- seul le modèle M1 permet d'expliquer abductivement d1 et d2,
- seul le modèle M2 permet d'expliquer abductivement d1 et d3.

Ainsi, dans l'exemple précédent, il n'est pas possible de préférer abductivement le modèle M1 au modèle M2 puisque tous deux sont de même simplicité *a priori* et tous deux expliquent des sous-ensembles de données maximaux pour l'inclusion ensembliste (et donc incomparables *a priori*). L'approche abductive pure nécessitera donc de rechercher des informations additionnelles pour faire un choix entre M1 et M2. Or, cette recherche est inutile : par déduction, on peut facilement montrer que M2 n'est pas le "bon modèle" puisqu'il est incompatible avec la donnée observée d2.

Notre approche peut donc être résumée comme suit :

Abduction Calcul des modèles expliquant les sous-ensembles les plus significatifs des données observées et sélection des meilleurs d'abord

Déduction Suppression des modèles incompatibles avec l'ensemble des données observées.

D'un point de vue informatique, elle peut être mise en œuvre en utilisant un système de maintien de vérité fondé sur des hypothèses (ATMS).

4. Résultats expérimentaux

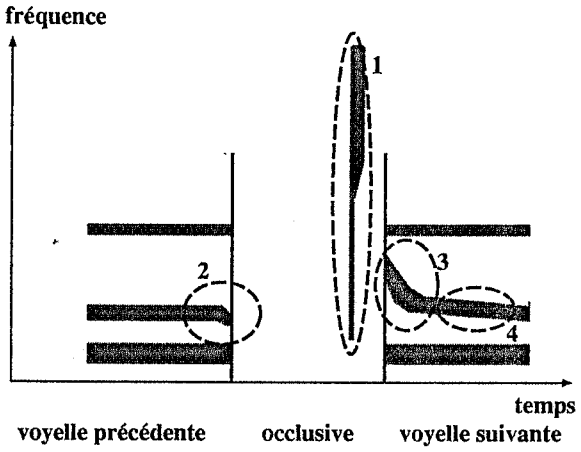
4.1 Méthodologie

Nous avons testé notre système sur un corpus d'occlusives sourdes (/p/, /t/, /k/) suivies par des voyelles postérieures. Ce corpus comprend deux parties. La première est constituée de 15 mots français monosyllabiques isolés, elle provient du corpus de mots isolés de BDSONS. Cette partie est enregistrée par deux locuteurs masculins et un locuteur féminin. La deuxième partie comprend 15 occurrences extraites d'un corpus de parole continue ("La bise et le soleil ... " de BDSONS), prononcées par des locuteurs masculins différents. Nous disposons donc de 60 occurrences au total.

Les données sont extraites de manière semi-automatique à l'aide de l'éditeur Snorri puis transférées vers le système. Aucune adaptation n'a été faite pour la locutrice. Nous travaillons actuellement à étendre ces résultats : d'une part nous allons faire plus de tests et d'autre part définir les stratégies correspondant aux occlusives suivies par des voyelles centrales et antérieures.

4.2 Déroulement d'un exemple

Nous décrivons le fonctionnement du système sur un exemple, ce fonctionnement est illustré par la figure 3.



1 "beau burst dental"
2 indice négatif dental : F2 ne se dirige pas du tout vers [1500, 2000]
3 indice négatif vélaire : F2 a une pente trop forte
4 l'occusive est suivie par une voyelle d'arrière

Figure 3 : illustration du raisonnement.

Après lecture des données, le système émet des hypothèses sur le contexte : ici il suppose que les contextes droit et gauche sont constitués par des voyelles d'arrière. Ensuite, il suppose tous les indices négatifs :

N1 (indice-négatif dentale direction-du-formant-F2-gauche)
N2 (indice-négatif vélaire F2-pente-trop-grande-droite).

L'hypothèse faite sur le contexte droit (voyelle d'arrière) déclenche la stratégie correspondante : l'indice de préférence

P (indice-préférence dentale beau-burst)

est trouvé, le système ne recherche alors aucun autre indice.

L'étape abductive conduit à supposer une dentale à partir de l'indice de préférence P, mais l'étape déductive élimine cette hypothèse car elle est contradictoire avec N1.

Pour l'instant, le processus s'arrête là. La contradiction peut provenir soit d'une erreur des détecteurs acoustiques, soit de la segmentation.

4.3 Résultats

Nous avons obtenu les résultats suivants :

Bonne reconnaissance	solution unique	parmi deux solutions
parole continue	11/15	13/15
mots isolés : locuteur	27/30	29/30
mots isolés : locutrice	8/15	11/15

Les principaux problèmes de notre système se manifestent lors de la reconnaissance de la locutrice et du /p/. Nous avons pu déterminer les causes principales d'erreur.
(i) une adaptation serait nécessaire pour la locutrice : les cas d'échec sont essentiellement dus à ce que F2 est trop élevé pour que le système puisse reconnaître une voyelle d'arrière ;
(ii) La moitié des cas d'échec provient d'une mauvaise détection du formant F2 : F2 n'a pas été trouvé a été confondu avec F3 par le suivi de formants. La voyelle suivant l'occlusive a été mal détectée, la stratégie utilisée n'était donc pas adaptée et n'a pas permis la reconnaissance de l'occlusive. De plus, les stratégies correspondant aux voyelles d'avant et centrales n'ont pas encore été définies et le module correcteur de formants n'a pas été utilisé pour ces tests. Ce module deviendra pertinent lorsque le système sera complété par une composante lexicale.
(iii) l'autre faiblesse importante du système est la mauvaise reconnaissance du son /p/. Nous n'avons pas assez d'occurrences de /p/ dans nos corpus pour mettre au point l'indice du barre d'explosion correspondant. Nous travaillons à élargir le corpus utilisable pour remédier à ce problème.
Finalement le système a reconnu correctement, comme unique solution, 46 occlusives sur 60, dont 40 grâce aux indices de préférence. Dans 7 autres cas, le système a proposé deux solutions dont la bonne. Pour les cas d'échec en reconnaissance de parole continue, le système n'a proposé aucune solution. En reconnaissace de mots isolés, il s'est trompé deux fois et n' a trouvé aucune solution dans les deux autres cas.

5. Conclusion

Nous avons présenté dans ce papier un nouveau modèle de reconnaissance automatique de la parole fondé sur l'utilisation de techniques de raisonnement hypothétique issues de l'intelligence artificielle. Ce modèle général a été dans un premier temps appliqué au décodage acoustico-phonétique des consonnes occlusives françaises. Les premiers résultats obtenus (86 % d'étiquetage correct sans aucune adaptation au locuteur) montrent que cette méthode permet une exploitation optimale des connaissances phonétiques.
Les principales contributions de ce travail sur le plan des modèles et des méthodes sont de trois ordres :
- un modèle théorique de raisonnement abductif/déductif adapté à l'indéterminisme de la parole,
- une méthode d'implantation pratique de ce modèle utilisant notre outil X-TRA et son système ATMS efficace,
- utilisation originale d'un mécanisme de retour en arrière "guidé par les échecs" pour résoudre les incohérences.

Nous travaillons actuellement au développement de ce projet dans plusieurs directions. D'abord nous étendons et validons la base de connaissances du niveau acoustico-phonétique, notamment pour affiner le mécanisme de retour en arrière. Ensuite nous étudions l'extension de notre modèle à d'autres niveaux de compréhension (lexical, syntaxique, sémantique) ainsi que les interactions entre niveaux. Nous comptons de ce fait parvenir à un modèle général unifié pour la gestion des hypothèses émises lors de la compréhension d'une phrase.

6 Références

F. Charpillet, P. Marquis, J.P. Haton [1991]. "X-TRA as a Toolbox for Truth Maintenance", Research report INRIA.

F. Charpillet, Y. Gong, J.P. Haton and D. Fohr [1991]. "AITRAS: A Real Time Expert System for Signal Understanding", World Congress on Expert System, Orlando.

R. De Mori, Yu F. Mong [1984]. "A System of Plans for Connected Speech Recognition", Proc. AAAI-1984, p. 92-95.

J. Doyle [1979]. "A Truth Maintenance System", *Artificial Intelligence* 12, pp. 231-272.

T. J. Edwards [1981]. "Multiple features analysis of intervocalic English plosives", JASA, Vol 69, No 2, pp. 535-547.

B. El Ayeb, P. Marquis and M. Rusinowitch [1990]. "Deductive / Abductive Diagnosis: the DA-Principles", Proc. *European Conference on Artificial Intelligence (ECAI–90)*, pp. 47-52, Stockholm.

D. Fohr and Y. Laprie [1989] "Snorri: an interactive tool for speech analysis", EUROSPEECH, Paris, pp. 669-672, volume 2.

R. Fox and J. R. Josephson : CV Experiment and Results. Technical Report, The Ohio State University, 1991.

J. de Kleer [1986]. "An Assumption-based TMS", *Artificial Intelligence* 28, pp. 127–162.

A. Komatsu, E. Oochira and A. Ichikawa [1991]. "Spontaneous Speech Understanding Based on Cooperative Problem Solving", IEICE Transactions 74, no. 7 July 1991, p. 1845-1853.

P. Marquis [1991]. "Towards Data Interpretation by Deduction and Abduction", Proc. *AAAI-91 Workshop on Abduction*, Anaheim (CA).

S. Nishioka, O. Kakusho and R. Mizoguchi [1991]. "A Generic Framework Based on ATMS for Speech Understanding System", IEICE Transactions 74, no. 7 July 1991, p. 1870-1880.

19es J.E.P. - Bruxelles, 19 au 22 mai 1992.

CODAGE PAR TRANSFORMEE ET SEGMENTATION AUTOMATIQUE : VERS UN CODEUR A DEBIT VARIABLE
Pour La Parole A Bande Elargie (0 à 7 kHz)

H. DIA, N. ACHAB et G. FENG

ICP URA CNRS N° 368, INPG/ENSERG Université Stendhal
BP 25X, 38040 GRENOBLE, CEDEX 9 FRANCE

Résumé

La transmission de la parole de haute qualité, c'est-à-dire à bande élargie (0 à 7 kHz), présente des applications potentielles dans les systèmes de communication telles que le visiophone et les téléconférences. Les codeurs existants peuvent satisfaire la qualité mais leurs débits sont encore trop élevés (>56 kbits/s). Nous présentons dans cette communication des résultats de recherche concernant la réalisation d'un codeur par transformée de 32 kbits/s pour la parole. Ce type de codage met à profit les caractéristiques psychoacoustiques du système autitif et a déja fait preuve d'efficacité pour le codage de la musique. Cependant l'élaboration du codeur pour la parole exige des traitements specifiques à ce signal. Ceux-ci constituent le point central de cette étude. Par ailleurs, nous présentons un algorithme de segmentation *on line*, basé sur la détection automatique de ruptures, qui sera intégré dans le codeur pour réaliser un système de codage à débit variable (debit moyen : 24 kbits/s).

I. INTRODUCTION

L'intégration des caractéristiques psycho-acoustiques du sytème auditif dans les systèmes de codage a récemment permis un progrès important dans le codage de haute qualité (Flanagan, 1991). Certes, dans les codeurs on pratique depuis un certain temps les mises en forme du bruit de quantification pour mieux le masquer (Atal et Schroeder 1979; Makhoul et Berouti 1979), et les premières propositions de codeurs tenant compte des propriétés auditives datent déjà de plus de dix ans (Schroeder et *al.*, 1979). Mais des progrès significatifs n'ont été enregistrés que récemment grâce au développement d'algorithmes de reconstitution parfaite du signal dans le domaine temporel "TDAC" (Princen et Bradley, 1986), et à l'apparition de processeurs de signal puissants. C'est dans ce contexte que le CNET a développé un système de codage de haute qualité pour la musique, basé sur le codage par transformée, la TDAC et le masquage fréquentiel, (Mahieux et Petit, 1990).

Le codage de parole à bande élargie (0 à 7 kHz) trouve des applications très importantes dans les télécommunications (visiophone, téléconférences, par exemple). La norme G722 du CCITT (ADPCM + sous-bandes) fournit une qualité satisfaisante mais le débit reste encore trop élevé (64 kbit/s). L'élaboration d'un codeur de la parole de même qualité mais avec un débit bien plus faible (24 à 32 kbits/s) constitue donc une tâche primordiale dans les applications citées ci-dessus. Notre recherche consiste d'abord à exploiter les expériences du codeur pour la musique, ce qui permettrait de réaliser un débit de 32 kbits/s. Un débit plus faible sera obtenu par la suite en utilisant un codage à débit variable. Une expérience concluante permettant une réduction importante de débit a déjà été réalisée sur un codeur CELP (Di Francesco, 1990).

L'adaptation d'un codeur pour la musique à la parole pour un débit de 32 kbits/s peut paraître simple : il suffirait de diviser la fréquence d'échantillonnage par deux. En effet, le codeur pour la musique fonctionne avec un débit de 64 kbits/s à la fréquence d'échantillonnage de 32 kHz, tandis qu'elle est de 16 kHz pour le codeur de la parole. Un essai de ce genre a été effectué et de bons résultats ont été obtenus (évaluations informelles), ce qui permet de justifier la démarche. Cependant, un retard de transmission intolérable, dû à l'emploi d'une taille de la transformée importante, empêche son application en pratique. Or, la diminution de cette taille entraîne des conséquences indésirables : réduction du taux de masquage et augmentation du débit. Le retard de transmission étant imposé, nous sommes donc amenés à utiliser une taille de transformée relativement faible et à optimiser le système de codage. Cette optimisation se justifie également par les grandes différences entre la musique et la parole (stationnarité, dynamique, etc...).

Dans cette communication, nous présentons d'abord les éléments essentiels du codeur, à savoir, la transformation en cosinus discrète modifiée, le masquage fréquentiel, le descripteur du bloc de transformée et l'allocation des bits. Ensuite, nous décrivons l'algorithme de segmentation basé sur la détection automatique de ruptures, destiné à la réalisation du codeur à débit variable.

II. MASQUAGE FREQUENTIEL

L'introduction du masquage fréquentiel dans un algorithme de codage par transformée a pour buts : 1). l'élimination des coefficients fréquentiels jugés inaudibles, 2). la mise en forme spectrale du bruit de codage. Cela exige l'emploi d'une transformation adéquate et la détermination d'une courbe de masquage optimale.

II.1. La TCDM

Un algorithme de codage par transformée efficace doit être associé à une transformation permettant une concentration de l'énergie sur le plus petit nombre de coefficients. Les transformations classiques telles que la TFD et la TCD ont des performances limitées en matière de concentration d'énergie par rapport à la Transformée en Cosinus Discrète Modifiée. Celle-ci, introduite par Princen et Bradley (1986), comporte une technique de suppression du repliment dans le domaine temporel, appelée TDAC (*Time Domain Aliasing Cancellation*). Cette transformation est définie de la manière suivante :

$$y(m,k) = \sum_{n=0}^{N-1} x(n).h(n)\cos(2\pi(2k+1)(2n+1)/(4N)+(2k+1)\pi/4)$$

avec k =0,...,N/2-1,

- N est la taille du bloc d'analyse, h(n) la fenêtre de pondération associée,
- m et k sont respectivement les indices temporel et fréquentiel.

La transformation inverse s'écrit :

$$x(m,n) = \sum_{k=0}^{N/2-1} y(m,k).f(n)\cos(2\pi(2k+1)(2n+1)/(4N)+(2k+1)\pi/4)$$

avec n =0,...,N-1 et f(n) est la fenêtre de synthèse.

La reconstruction du signal du bloc m est obtenue par l'application de la technique d'*overlap-add*. Pour une reconstruction parfaite du signal en l'absence du bruit de codage, il faut que h(n) = f(n). La fenêtre utilisée est celle de Hanning.

Le calcul des transformations directe et inverse par les formules citées ci-dessus ne permettrait pas un fonctionnement de l'algorithme en temps réel. Il existe un algorithme de calcul rapide de la TCDM, développé par Duhamel & *al.* (1991).

II.2. Courbe de masquage

La détermination de la courbe de masquage s'effectue sous l'hypothèse de linéarité suivante : elle est la somme des formes de masquage élémentaires $S_k(m,i)$ correspondant à chaque coefficient de transformée. La contribution du coefficient y(m,k) au masquage des autres coefficients y(m,i) avec $i \neq k$ s'écrit :

$$S_k(m,i) = B(\mu_i - \mu_k).a_0(k).y^2(m,k)$$

Le calcul de la courbe de masquage revient à convoluer l'ensemble des coefficients au carré du bloc de transformée TCDM, pondéré par le facteur de transmission de l'oreille $a_0(k)$, par la fonction d'étalement spectral de la membrane basilaire $B(\mu)$.

$$S(m,i) = \phi . \sum_{k=0}^{N/2-1} B(\mu_i - \mu_k).a_0(k).y^2(m,k)$$

Le seuil de masquage $\phi = 0.001$ (-30 dB) représente le seuil nécessaire déterminé par Zwicker et Feldtkeller (1981) pour qu'un son pur masque un autre lorsqu'ils sont dans la même bande critique. $B(\mu)$, la fonction d'étalement spectral de la membrane basilaire, est composée de trois parties :

- la partie centrale appelée sonie de cœur où $B(\mu) = 1$ pour $-1/2$ Bark $< \mu < 1/2$ Bark ;
- les deux autres parties représentant respectivement les sonies de flanc vers les fréquences supérieures avec une pente de -10 dB par Bark et vers les fréquences inférieures avec une pente de -27 dB par Bark.

En définitive, un coefficient y(m,k) sera masqué si $a_0(k).y^2(m,k) < S(m,k)$

La détermination de la courbe de masquage globale par la formule donnée ci-dessus nécessite une charge de calcul importante. Mahieux & Petit (1990) ont développé un algorithme de calcul rapide en constatant que les formes de masquage voisines présentent peu de variations. On peut donc utiliser la même forme de masquage pour un certain nombre de coefficients groupés. Cela revient à découper le bloc de TCDM en sous-bandes de largeur inégale.

II.3. Détermination du seuil de masquage optimal

Le but du masquage est d'obtenir un maximum de coefficients masqués sans dégrader la qualité du signal reconstruit. Il nous semble que pour des sons complexes comme la parole on peut baisser (en valeur

absolue) la valeur de ø par rapport à celle donnée par Zwicker et Feldtkeller (1981) pour des sons purs. Afin d'obtenir un bon critère pour le choix de ø, nous avons évalué son influence sur les signaux reconstruits à l'aide d'un système d'analyse-synthèse comprenant les procédures suivantes :

- calcul de la TCDM directe avec une fenêtre de 512 points ;
- calcul de la courbe de masquage et suppression de coefficients masqués ;
- calcul de la TCDM inverse et synthèse.

Les signaux obtenus avec ce système d'analyse-synthèse sont comparés aux originaux pour différentes valeurs de ø. Trois types de signal de parole sont utilisés :

- parole continue (12 phrases),
- une voix chantée,
- une voix chantée composée que de voyelles.

Nous avons obtenu par cette expérience une valeur de ø = -18 dB, en dessous de laquelle les signaux reconstruits ne présentent pas de dégradations perceptibles.

L'analyse du pourcentage de raies masquées en fonction de ø (figure 1) montre qu'on ne peut masquer, pour des signaux de parole continue, plus de 63% de coefficients (|ø| = 18 dB) pour une fenêtre d'analyse de 512 points.

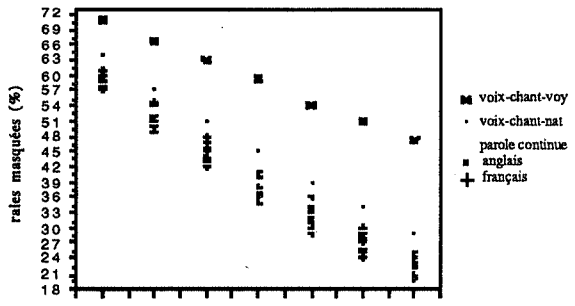


Fig.1. Pourcentage de raies masquées en fonction du seuil |ø| exprimé en dB.

On peut remarquer que les taux de raies masquées pour la voix chantée composée que de voyelles sont nettement supérieurs à ceux obtenus pour la parole continue. Il semble clair que les sons non voisés sont à l'origine de ces pourcentages réduits. En effet, la structure des coefficients de la TCDM étant proche de celle d'une TFD classique, elle présente une forme harmonique pour des sons voisés, ce qui permet un meilleur masquage (figure 2a). Pour des sons non voisés, l'absence de cette structure harmonique entraîne un taux de masquage relativement faible (figure 2b).

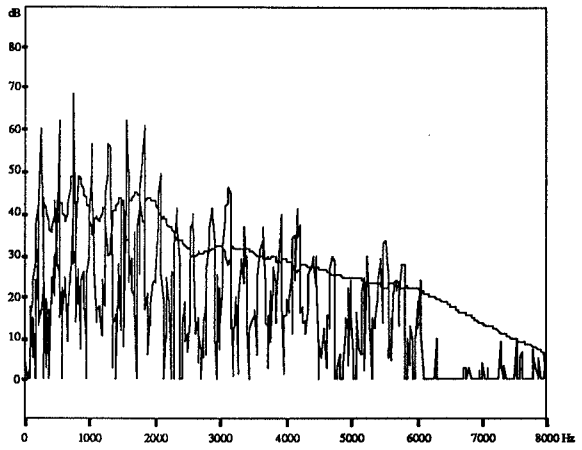


Fig.2a. Coefficients de la TCDM et courbe de masquage associée pour une voyelle /a/. Taille de la transformée : 1024 points et ø = -20 dB. Nombre de coefficients masqués : 410 (sur 512).

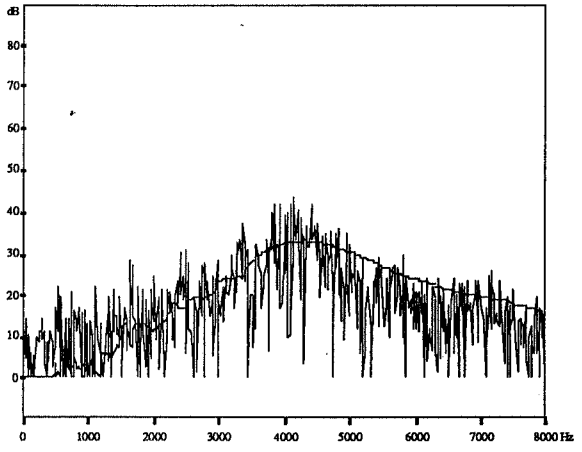


Fig.2b. Coefficients de la TCDM et courbe de masquage associée pour une consonne /s/. Taille de la transformée : 1024 points et ø = -20 dB. Nombre de coefficients masqués : 305 (sur 512).

III. DESCRIPTEUR

Pour obtenir la même allocation des bits tant au codeur et au décodeur, il faut une information auxiliaire (descripteur) très précise sur la représentation fréquentielle du bloc de transformée. Le descripteur correspond à l'écart-type de l'énergie moyenne des coefficients de TCDM calculée pour chaque sous-bande.

- ***Prédiction***

Pour mieux exploiter la corrélation inter-blocs ou intra-bloc, deux types de prédiction sont testés et le choix definitif sera effectué selon leur efficacité :

1) la prédiction temporelle : soit $\sigma(m,j)$ le descripteur de la sous-bande j, et $\sigma'(m-1,j)$ la valeur quantifiée du bloc précédent, l'erreur de prédiction s'écrit :

$e(m,j) = \log(\sigma(m,j)) - \log(\sigma'(m-1,j))$

avec $j = 0,...,N_{sb}-1$ ($N_{sb} = 32$ sous-bandes).

2) la prédiction fréquentielle : l'erreur de prédiction fréquentielle s'écrit :

$e(m,j) = \log(\sigma(m,j)) - \log(\sigma'(m,j-1))$

avec $j = 1,...,N_{sb}-1$.

Le descripteur $\sigma(m,0)$ est quantifié séparément sur 7bits.

- ***Statistiques et optimisation***

L'étude statistique des erreurs de prédiction temporelle et fréquentielle montre que l'emploi d'un quantificateur uniforme de dynamique 80 dB avec un pas de 5 dB est satisfaisant. Comme le masquage n'est pas uniforme d'un bloc à un autre et d'une sous-bande à une autre, la dynamique des erreurs de prédiction est variable. L'exploitation de cette variabilité se traduit par l'utilisation du codage de Huffman.

Après le calcul du nombre de bits par bloc, le type de prédiction qui donne le moins de bits sera retenu et un bit supplémentaire sera utilisé pour le spécifier au décodeur. Nous avons déterminé les mots de code de Huffman pour les deux prédicteurs, en respectant ce critère de sélection, sans nous préocupper de la nature (voisé/non voisé) du signal.

Notre analyse statistique montre que l'on peut obtenir une réduction du nombre de bits moyen par sous-bande en séparant le signal de parole en deux classes : voisée et non voisée (voir tab. 1).

Prédiction	Temporelle	Fréquentielle
Sans classement	2.60	2.68
Voisé/non voisé	2.58	2.62

Tab.1 Nombre de bits moyen par sous-bande pour coder les erreurs de prédiction.

IV. ALLOCATION DES BITS

La procédure de l'allocation des bits doit répartir le débit entre les différents coefficients d'un bloc de TCDM, de telle sorte que le bruit de quantification soit masqué et que le nombre de bits disponible ne soit dépassé. Ceci revient à minimiser, pour chaque bloc m, l'expression :

$$X(m) = \sum_{k=0}^{N/2-1} \sigma_q^2(m,k)/S(m,k)$$

avec $\sigma_q^2(m,k)$ la variance du bruit de quantification du coefficient y(m,k), S(m,k) le seuil de masquage correspondant.

La minimisation de X(m) s'effectue en respectant la contrainte $\Sigma\ R(m,k) = R_0(m)$, R(m,k) étant le nombre de bits alloués à y(m,k), et $R_0(m)$ le nombre total de bits à répartir. Jayant et Noll (1984) ont proposé, pour trouver le nombre de bits pour y(m,k), de résoudre par procédure itérative l'équation :

$$R(m,k) = 1/2.(\log_2(\sigma^2(m,k)/S(m,k))+\lambda(m)$$

avec $\lambda(m)$ constante, $\sigma^2(m,k)$ la variance de y(m,k).

Comme l'allocation des bits doit être la même au codeur comme au décodeur, on calcule R(m,j) au lieu de R(m,k). Donc chaque coefficient y(m,k) de la sous-bande j recevra le même nombre de bits. $\sigma^2(m,k)$ sera remplacé par $\sigma'^2(m,j)$, et S(m,k) par $S_1(m,j)$, le nouveau seuil de masquage calculé à partir de $\sigma'^2(m,j)$. Mahieux et Petit (1990) ont proposé de résoudre par itérations l'équation suivante :

$$R(m,j) = 1/2.(\log_2(\sigma'^2(m,j)/S_1(m,j))+\lambda(m)$$

Si à la fin de la procédure itérative la contrainte relative au nombre de bits n'est pas respectée, on ajoute ou retranche des bits selon la méthode de Perkins et Lookabaugh (1989). Ensuite, les coefficients sont quantifiés à l'aide d'un quantificateur non uniforme.

V. SEGMENTATION

Un système de codage à débit variable consiste à faire varier la longueur des trames de façon à permettre l'exploitation systématique des redondances et optimiser les paramètres du codage sur les segments relativement stationnaires du signal de parole. Pour la réalisation d'un tel système, il faut d'abord segmenter le signal ; aussi doit-on donc disposer d'une détection séquentielle et automatique de frontières (ruptures). Cette segmentation doit être simple, efficace, et surtout "on-line".

V.1. Principe

Le signal de parole peut être considéré comme une juxtaposition de zones stationnaires. Pour détecter séquentiellement des changements dans les caractéristiques spectrales du signal, ce dernier est observé sur deux échelles de temps (fenêtres) : une observation à long terme et une autre à court terme. Ces deux tranches sont modélisées et, à chaque modèle est associée une densité de probabilité. Un test statistique est ensuite utilisé pour mesurer la distance entre ces deux modèles. La détection étant séquentielle cette mesure s'effectue à l'acquisition de chaque nouvel échantillon. Cette distance reste faible si l'on se trouve en zone stationnaire et devient importante lorsqu'un changement dans la nature du signal survient. Une

procédure de détection adéquate basée sur cette distance permet de détecter ces changements (appelés souvent "ruptures").

V.2. Test statistique

Le test statistique est construit sur les hypothèses suivantes : les observations avant rupture suivent le modèle long terme (hypothèse H_0) et, après rupture, le modèle court terme (hypothèse H_1). Ce test, introduit pour les séries temporelles [Basseville et Benveniste, 1983], est du type somme cumulée (cusum). Il est obtenu à partir du rapport de vraisemblance *a posteriori* entre les deux distributions corrigé par la divergence de Kullback [Kullback, 1959]. Dans le cas simple de la modélisation autorégressive et des résidus gaussiens, la statistique w_n , incrément de la somme cumulée, est donnée par l'expression suivante :

$$w_n = \frac{1}{2}\left[\frac{2\, e_n^1 e_n^0}{\sigma_1^2} - \left[1 + \frac{\sigma_0^2}{\sigma_1^2}\right]\left(\frac{e_n^0}{\sigma_0}\right)^2 + \left[1 - \frac{\sigma_0^2}{\sigma_1^2}\right]\right]$$

avec e_n^i (i=0,1) résidus des modèles long et court terme, et σ_i^2 la variance des résidus correspondants.

Théoriquement la somme cumulée de ces incréments oscille autour de zéro dans les zones stationnaires et chute à chaque changement spectral. Pour réduire le retard à la détection et obtenir une meilleure estimation de l'instant de rupture, un biais fixé *a priori* est ajouté à chaque incrément (test cusum de Hinkley, 1971).

V.3. Modélisation

La modélisation du signal joue un rôle très important parce qu'elle permet aux signaux résiduels d'approcher les hypothèses sur lesquelles la statistique du test est établie. Pour le signal de parole, les modèles couramment utilisés sont linéaires, autorégressifs. Quoique simples, ces derniers sont représentatifs car ils reflètent correctement les caractéristiques du conduit vocal.

Dans l'algorithme de segmentation développé pour la reconnaissance de la parole [André-Obrecht, 1985,1988], le modèle long terme est identifié par l'algorithme de Burg sur une fenêtre croissante et le modèle court terme par l'autocorrélation sur une fenêtre glissante. Un autre algorithme développé pour le codage à débit variable [Di Francesco, 1990] utilise pour l'identification du modèle long terme l'algorithme de Burg avec coefficient d'oubli et pour le modèle court terme la méthode de covariance en treillis.

Il faut souligner que le test utilisé dans les deux algorithmes mentionnés ci-dessus présente un caractère de dissymétrie : le test peut réagir lors de la transition entre deux zones stationnaires consécutives, alors qu'il peut rester insensible lorsque la position de ces deux zones est inversée. Cela engendre des omissions de détection. Pour pallier à cet inconvénient, une solution consiste à segmenter le signal dans les deux sens [André-Obrecht, 1988] mais elle n'est pas envisageable pour le codage.

Dans une étude précédente, Feng et *al.*, (1991) ont établi le lien entre la dissymétrie de ce test et la modélisation du signal. En effet une modélisation inadéquate peut accentuer la dissymétrie tandis qu'elle peut être atténuée par l'utilisation de modèles adaptatifs. Après une étude comparative, l'algorithme de Burg séquentiel avec coefficient d'oubli [Benveniste, 1983] et la méthode *prewindow* adaptative [Friedlander, 1982] sont retenus dans l'algorithme de segmentation.

Les équations de ces modèles adaptatifs sont :

- *Algorithme de Burg adaptatif*

pour p = 1, min(ordre,T)

$$k_{p,T} = 2\, c_{p,T} / (R^e_{p,T} + R^e_{p,T})$$
$$e_{p+1,T} = e_{p,T} - k_{p,T}\, r_{p,T-1}$$
$$r_{p+1,T} = r_{p,T-1} - k_{p,T}\, e_{p,T}$$
$$R^e_{p+1,T} = \lambda_0\, R^e_{p+1,T-1} + (1 - \lambda_0)\, e^2_{p+1,T}$$
$$R^r_{p+1,T} = \lambda_0\, R^r_{p+1,T-1} + (1 - \lambda_0)\, r^2_{p+1,T}$$

- *Algorithme Prewindow*

pour p = 0, min(ordre,T)-1

$$c_{p+1,T} = \lambda_1\, c_{p+1,T-1} + e_{p,T}\, r_{p,T-1} / \gamma^c_{p-1,T-1}$$
$$\gamma^c_{p,T} = \gamma^c_{p-1,T} - r^2_{p,T} / R^r_{p,T}$$
$$k^r_{p+1,T} = c_{p+1,T} / R^r_{p,T-1}$$
$$R^e_{p+1,T} = R^e_{p,T} - k^r_{p+1,T}\, c_{p+1,T}$$
$$k^e_{p+1,T} = c_{p+1,T} / R^r_{p,T}$$
$$r_{p+1,T} = r_{p,T-1} - k^e_{p+1,T}\, e_{p,T}$$
$$R^r_{p+1,T} = R^r_{p,T-1} - k^e_{p+1,T}\, c_{p+1,T}$$

Remarque : $1/(1 - \gamma_{p,T})$ est le gain adaptatif qui permet de suivre rapidement les variations de la statistique du second ordre du signal.

V.4. Détection de ruptures

L'utilisation des modèles adaptatifs permet d'éviter la réinitialisation après chaque rutpure. Par conséquent la procédure de détection peut rester relativement simple. Nous n'utiliserons pas la technique de détection classique basée sur le maximum local puisqu'il n'y a pas de réinitialisation.

Nous avons constaté qu'aux variations spectrales importantes du signal correspondent toujours des changements de pente dans la statistique. Ce sont donc ces changements que nous devons détecter. Pour cela, nous linéarisons d'abord la statistique afin de masquer ses variations microscopiques. Ensuite, nous comparons les pentes de deux segments successifs : une rutpure sera décidée si la différence dépasse un seuil.

V.5. Résultats

L'algorithme de segmentation a été évalué sur plusieurs corpus contenant des logatomes CVCV et de la parole continue échantillonnés à 16 kHz et codés sur 16 bits.

Les résultats de la segmentation sont satisfaisants : les changements spectraux importants sont correctement détectés.

Il faut mentioner que la détection de rutpures est toujours d'autant plus difficile que le spectre varie lentement. Mais pour le codage la non détection des changements spectraux lents n'est pas un inconvénient majeur. En effet, le signal peut dans ce cas être considéré comme stationnaire.

VI. CONCLUSION

Nous avons présenté dans cette communication des résultats de recherche concernant l'élaboration d'un codeur par transformée de la parole à bande élargie (0-7kHz). Cette étude permet de transformer un codeur de même type pour la musique (64 kbit/s, 32 kHz), développé par le CNET, en un codeur pour la parole. Nous avons réalisé plusieurs étapes d'optimisation nécessaires pour cette transformation : détermination du masquage optimal, analyse statistique pour le codage du descripteur et une étude prélimiaire sur l'allocation des bits. Les résultats de cette étude permettront la réalisation définitive du codeur de 32 kbits/s.

Nous avons également présenté un algorithme de segmentation *on line* de la parole. Basé sur un test statistique et sur des modèles adaptatifs, il présente les avantages d'être efficace et robuste.

Cet algorithme de segmentation sera intégré dans le codeur de 32 kbits/s pour constituer un système de codage à débit variable qui permettra de réaliser un débit moyen de 24 kbits/s.

REMERCIEMENTS

Cette étude a été financée pour l'essenciel par le département TSS/CMC du CNET Lannion A (P. Comberscure, Convention No 90 7B 051). Nous remercions Y. Mahieux pour ses conseils et pour les discussions fructueuses que nous avons eues avec lui.

BIBLIOGRAPHIE

André-Obrecht R. (1985), Segmentation automatique du signal de parole. Thèse de 3ème cycle, Univ. de Rennes I.

André-Obrecht R. (1988), A new statistical approach for the automatic segmentation of continuous speech signals. IEEE Trans. ASSP, vol. 36, n°1, 29-40.

Atal B.S. & Schroeder M.R. (1979), Predictive coding of speech signals and subjective error criteria. IEEE Trans. ASSP, vol. 27, n°3, 247-254.

Basseville M. & Benveniste A. (1983), Sequential detection of abrupt changes in spectral characteristics of digital signals. IEEE Trans. Inform. Theory, vol. 29, n°5, 708-723.

Benveniste A. (1983), Algorithmes simples d'estimation en treillis pour les séries longues. in "Outils et modèles mathématiques pour l'automatique, l'analyse des systèmes et le traitement du signal", CNRS, vol. 2, 309-330.

Di Francesco R.J. (1990), Real time speech segmentation using pitch and convexity jump models : application to variable speech rate coding. IEEE Trans. ASSP, vol. 38, n°5, 741-748.

Duhamel P., Mahieux Y. & Petit P. (1991), A fast algorithm for the implementation of filter banks base on time domain aliasing cancellation, Proc. of ICASSP, Toronto,vol. 3, 2209-2212.

Feng G., Achab N., & Combescure P. (1991), On-Line speech segmentation using adaptive models : application to variable speech coding. Eurospeech 1991, Genova, Italy, vol. 2, 705-708.

Flanagan J.L. (1991), Speech technology and computing: a unique partnership. Eurospeech 1991, Genova, Italy, vol. "opening session", 7-22.

Friedlander B. (1982), Lattice filters for adaptive processing. Proc. IEEE, vol 70, n°8, 829-867.

Hinkley D.V. (1971), Inference about the change-point from cumulative sum tests. Biometrika, vol 58, n°3, 509-523.

Jayant N. & Noll P. (1984), "Digital coding of waveforms", Prentice Hall Signal Processing Series.

Kullback S. (1959), "Information theory and statistics". John Willey & sons Inc., New York.

Mahieux Y. & Petit J. (1990), Transform coding of audio signals at 64 kbits/s, Globecom'90, San Diego, 518-522.

Makhoul J. & Berouti M. (1979), Adaptive noise spectral shaping and entropy coding in predictive coding of speech. IEEE Trans. ASSP, vol. 27, n°1, 63-73.

Perkins M. & Lookabaugh T. (1989), "A psychophysically justified bit allocation algorithm", Proc of ICASSP, Glasgow, vol. 3, 1815-1818.

Princen J.P., Bradley A. (1986), Analysis/synthesis filter bank design based on time domain aliasing cancellation, IEEE Trans. ASSP., vol. 34, n°5, 1153-1161.

Schroeder M.R., Atal B.S., & Hall J. (1979), Optimizing digital speech coders by masking properties of the human ear. JASA, vol 66, n°6, 1647-1652.

Zwicker E., Feldtkeller R. (1981), "Psychoacoustique, l'oreille réceptrice d'information", Ed. Masson.

19es J.E.P. - Bruxelles, 19 au 22 mai 1992.

A LA RECHERCHE DE L'ESPACE DISTAL DE CONTRÔLE EN PAROLE : LA PISTE DES TUBES LABIAUX.

C. SAVARIAUX, P. PERRIER & L.J. BOË

Institut de la Communication Parlée - U.R.A. CNRS N° 368
INPG & Université Stendhal, 46 Av. Félix Viallet
38031 Grenoble Cédex - France

Résumé

Nous proposons ici les premiers résultats d'une étude expérimentale visant à caractériser l'espace distal de contrôle en parole. L'expérience consiste à étudier les stratégies développées par un locuteur français (après apprentissage), lorsque ses conditions naturelles de production vocalique sont perturbées. La perturbation porte sur l'ouverture labiale, qui est alors imposée par la section d'un tube labial qui empêche la réalisation de la forme naturelle du conduit vocal. Les résultats, analysés tant sur le plan articulatoire qu'acoustique, montrent que le locuteur n'a pas produit un signal perceptivement correct, et ceci quelle que soit la taille du tube labial. Pourtant, dans chaque cas, le locuteur a réagi à la perturbation en modifiant la forme de sa langue, mais sans pour cela altérer la position et la taille de la constriction. Sur quels critères (acoustiques ou articulatoires) s'est-il basé pour procéder à cette "compensation" ? C'est à cette délicate question que nous tentons, en conclusion, d'apporter des éléments de réponse.

1. INTRODUCTION

Le système périphérique de production de la parole se caractérise, comme tous les systèmes biologiques, par son adaptabilité. Rappelons en deux manifestations essentielles :

- selon le contexte phonétique, un même phonème peut avoir des réalisations différentes, articulatoirement ou acoustiquement ; il s'agit du résultat des mécanismes de **coarticulation**, reflétant la négociation entre la nécessité d'atteindre des cibles perceptives et l'exigence d'optimisation gestuelle (minimisation de l'"effort").
- le fumeur de pipe réussit à s'exprimer de manière compréhensible tout en maintenant sa pipe entre ses dents ; ce paradigme classique met en évidence la capacité de chacun à produire un même son pour des positions différentes de ses articulateurs (mécanismes de **compensation** des perturbations).

Poser le problème de la gestion de cette adaptabilité, revient à poser, de manière très générale, celui de l'**équifinalité** de deux gestes en parole ; et donc celui de la latitude dont on dispose articulatoirement dans la réalisation d'un même objectif acoustico-perceptif. Nous proposons ici d'apporter quelques éléments à ce débat avec, en toile de fond, une question clé : **quels sont, dans les espaces acoustique et articulatoire, les objectifs autour desquels se gère la variabilité des gestes de la parole ?**

L'expérience de référence en la matière est celle des *bite-blocks* de Gay et al. (1981). Les auteurs ont analysé, lors la production de voyelles et pour différents locuteurs, les modifications de l'articulation (observées dans le plan sagittal du conduit vocal par rayons X) induites par l'insertion d'une "cale" (le *bite-block*), qui, une fois serré entre les dents, bloque la position de la mandibule. Contrairement à ce que l'on pouvait pronostiquer à partir des prédictions des modèles acoustiques (cf. par exemple Mermelstein, 1967 ; Schroeder, 1967), on observe que, dans tous les cas, en présence de *bite-blocks*, les locuteurs se sont efforcés de maintenir la position et la taille de la constriction dans le conduit vocal quasiment identiques à celles de la production sans contrainte ; les auteurs de l'expérience en concluaient que : "*The target of a vowel is coded neurophysiologically in terms of area-function related information and is specified with respect to the acoustically most significant area-function features, the point of constriction along the length of the tract*" . Cependant, pour maintenir constante la géométrie du conduit vocal dans ses zones "cruciales", les locuteurs ont dû modifier le positionnement naturel de leurs articulateurs : il s'agit de **compensation articulatoire**. L'importance, semble-t-il primordiale, de la zone de constriction dans le conduit vocal avait par ailleurs été déjà soulignée par Wood (1979) qui, en se fondant sur l'analyse radiographique de voyelles de l'anglais et de l'arabe, montrait que le lieu d'articulation des voyelles ne pouvait être placé que dans 4 zones bien distinctes du conduit vocal, et, plus important encore, qu'à chacune de ces zones de constriction ne pouvait être

associée qu'une catégorie bien spécifique de voyelles (Wood, 1979). Ces conclusions ont récemment été confirmées par Boë et al. (1992), grâce à l'exploitation en extension d'un modèle anthropomorphique du conduit vocal, qui, à partir de paramètres articulatoires du type mandibule, langue et lèvres, génère un signal vocalique (Maeda, 1979) : à l'exception de la voyelle centrale [œ], dont la constriction est très peu marquée, toutes les voyelles (caractérisées par leurs trois premiers formants) présentent une seule zone de constriction. De plus, cette dernière étude montrait aussi clairement que pour les voyelles arrondies (du type [u] ou [y]) de fortes contraintes s'exercent sur les lèvres : l'aire labiale doit impérativement rester très faible.

Ces différentes études menées à l'interface des domaines articulatoire et acoustique pondèrent donc nettement les prédictions faites à partir des seuls modèles acoustiques et les précisent : parmi toutes les possibilités théoriques de compensation, les locuteurs n'ont recours, en présence de perturbation, qu'à celles qui n'affectent pas les "paramètres cruciaux" de la géométrie du conduit vocal (géométrie de la constriction et de l'orifice labial). Faut-il y voir la preuve que le contrôle en parole a pour finalité, non pas une caractéristique acoustique du signal de parole, mais une forme spécifique du conduit vocal, décrite par ces paramètres cruciaux ? Certes nous avons là des indices intéressants qui tendent à corroborer cette hypothèse, mais aucun des résultats n'est réellement décisif. En effet les tracés de Wood sont obtenus sur de la parole naturelle, tout comme le modèle articulatoire de Maeda ; la seule conclusion que l'on peut formuler est donc qu'en parole naturelle, la production d'une voyelle donnée correspond à une seule position de la constriction et/ou à une forme précise des lèvres. Mais cette correspondance biunivoque est-elle l'émanation d'un **objectif strictement articulatoire**, ou est-ce la conséquence d'une optimisation gestuelle qui, après la longue période d'apprentissage de la parole, a déterminé les stratégies articulatoires les mieux appropriées à la réalisation d'un **objectif acoustique** donné ? Cette équivoque ne peut être levée que dans la mesure où l'on place le locuteur dans une situation telle que les associations spontanées liées à l'apprentissage ne soient plus possibles. De ce point de vue, l'expérience des *bite-blocks* n'est pas complètement satisfaisante car elle n'empêche pas réellement les locuteurs d'atteindre la forme du conduit vocal associée, par apprentissage, à la voyelle étudiée.

C'est à cet objectif que nous nous sommes attachés pour la définition de notre protocole expérimental à l'aide de tubes labiaux.

2. L'EXPÉRIENCE DES TUBES LABIAUX

2.1. Protocole expérimental

Perturber la forme du conduit vocal dans ses zones cruciales, revient donc à affecter soit le lieu d'articulation, soit les lèvres. Du point de vue pratique, il est bien clair qu'il est plus aisé de perturber l'aire labiale. C'est donc ce que nous avons fait à l'aide de petits tubes creux en plexiglas (les **tubes labiaux**) dont le diamètre est tel que la forme cible des lèvres ne peut pas être atteinte. L'expérience n'est évidemment pertinente que pour des voyelles sensibles aux lèvres, les voyelles arrondies. Parmi celles-ci, le [u] est particulièrement intéressant, car il possède un lieu de constriction palato-vélaire, et donc relativement central dans le conduit vocal ; on peut donc envisager que cette voyelle est celle qui offre le plus de possibilités de compensations avant/arrière.

Notons que le modèle de Maeda prédit effectivement deux lieux possibles pour obtenir les deux premiers formants caractéristiques du [u] : l'un, antérieur, est associé à un contrôle précis des lèvres (c'est le standard du [u] français), tandis que l'autre plus postérieur, laisse plus de latitude sur le contrôle labial (voyelle du type [U] anglais) ; ces deux réalisations vocaliques se distinguent sur la valeur du troisième formant (Boë et al., 1992).

L'aire standard moyenne pour le [u] prononcé par notre locuteur a été déterminée après l'analyse et l'interprétation d'images vidéos des lèvres (Lallouache & Worley, 1988 ; Lallouache, 1991) dans différents contextes phonétiques. Nous en avons déduit une valeur moyenne de 4 mm pour la hauteur des lèvres. Nous avons donc utilisé des tubes labiaux de 2 et 30 mm de diamètre extérieur, ce qui correspond respectivement à une fermeture et à une nette ouverture des lèvres ; chaque tube a une longueur de 25 mm.

Les mesures sont effectuées sur un locuteur français de sexe masculin, après une période d'apprentissage non contrôlée (une journée avant les prises de vue). La consigne pour le locuteur consistait à s'efforcer de reproduire la voyelle [u] tout en maintenant ses lèvres serrées sur le tube. Le déclenchement des rayons X n'est intervenu qu'une fois le son stabilisé pendant au moins 2 secondes, et ceci avec les deux tubes labiaux. Une radiographie a préalablement été effectuée pour la voyelle [u] en contexte normal afin d'avoir la forme de référence.

Les clichés par rayons X ont été réalisés au Centre Hospitalier Universitaire de Grenoble. Ils représentent une vue latérale du conduit vocal dans le plan sagittal, toute la tête étant englobée à l'exclusion de la glotte. Parallèlement un enregistrement du signal acoustique a été effectué sur cassette vidéo large bande (standard Betamax), pour permettre l'évaluation ultérieure de la réalisation à partir des paramètres formantiques.

A partir des tracés radiographiques, nous avons réalisé la détection et l'acquisition des contours sagittaux du conduit vocal partant du haut du larynx jusqu'à l'extrémité des lèvres. Le contour externe de la mandibule a aussi été dessiné. Les paramètres formantiques (trois premiers formants) sont extraits par la méthode du cepstre.

2.2. Résultats et interprétations

Chaque tracé obtenu avec un tube labial a été superposé au tracé obtenu en parole naturelle. Cette superposition a été réalisée en prenant comme référence, pour une analyse globale, la partie frontale de la boite crânienne et l'incisive supérieure qui correspondent à deux points de référence fixes chez le locuteur, et, pour l'étude de la modification du geste lingual, la mandibule.

Pour interpréter acoustiquement les modifications de la géométrie du conduit vocal, nous avons eu recours à un modèle acoustique classique : le modèle à quatre tubes de Fant (Fant, 1960) basé sur le principe de la modélisation acoustique par ondes planes stationnaires. Ce modèle, bien que simple dans son principe de modélisation, permet une bonne prédiction des formants à partir de la géométrie du conduit vocal. Fant a ainsi pu proposer, pour deux valeurs de l'aire à la constriction et pour cinq valeurs de l'aire labiale, des nomogrammes montrant l'évolution des cinq premiers formants avec la variation de la position avant/arrière de la constriction dans le conduit vocal. On peut ainsi y repérer le patron formantique propre à la voyelle [u] et en déduire les modes de résonance du conduit vocal correspondants. On peut alors associer à chacun des formants du [u] les modes de résonances suivants : F1 est principalement lié à la résonance basse du Helmholtz formé par la cavité arrière et la constriction ; F2 est majoritairement le résultat de la résonance basse du Helmholtz formé par la cavité avant et les lèvres ; F3 est la fréquence de résonance demi-onde de la cavité arrière (pour une discussion sur ces problèmes d'affiliation majoritaire, cf. en particulier Badin et al., 1990).

Pour la voyelle [u], à partir de ces nomogrammes, la constriction étant centrée approximativement à 10 cm de la glotte, on peut prédire l'ordre de grandeur des trois premiers formants : 280 Hz pour F1, 700 Hz pour F2, et 2200 Hz pour F3. Ces valeurs sont en effet en bonne correspondance avec les mesures effectuées sur le signal acoustique de notre locuteur : F1 = 270 Hz, F2 = 562 Hz, F3 = 2197 Hz.

2.2.1. Cas du tube labial de 2 mm

• *Analyse de la géométrie du conduit vocal* : Entre le contour sagittal du [u] avec le tube labial de 2 mm et celui du [u] en contexte normal (figure 1) on remarque essentiellement une modification de la position de la langue : celle-ci se trouve plus avancée et plus haute dans sa partie antérieure ; il s'ensuit un léger déplacement de la constriction vers l'avant. Le tube labial de faible diamètre induit, outre la modification recherchée de l'orifice labial, une rétraction des lèvres. On note enfin un abaissement du vélum.

• *Prédictions des conséquences acoustiques* : Selon le principe de Helmholtz, la fermeture des lèvres devrait induire une diminution de F2. Cependant la diminution conjointe du volume de la cavité avant, due à l'avancée de la langue, tend à compenser l'effet acoustique de cette perturbation labiale : F2 devrait donc rester sensiblement constant. L'avancée de la langue augmente la longueur (et donc le volume) de la cavité arrière, ce qui aura deux conséquences acoustiques : (1) une diminution de F3 (résonance demi-onde de la cavité arrière) et (2) une diminution de F1 (Helmholtz formé par la cavité arrière et la constriction) ; cependant compte tenu de la présence d'une masse d'énergie basse fréquence imputable aux vibrations des parois du conduit vocal, il est vraisemblable que la diminution de F1 sera masquée.

Enfin, l'abaissement du vélum crée un couplage acoustique entre les cavités buccale et nasales, et deux fréquences de résonance doivent apparaître autour de 250 et 1000 Hz (Feng et al., 1986 ; Castelli et al., 1989).

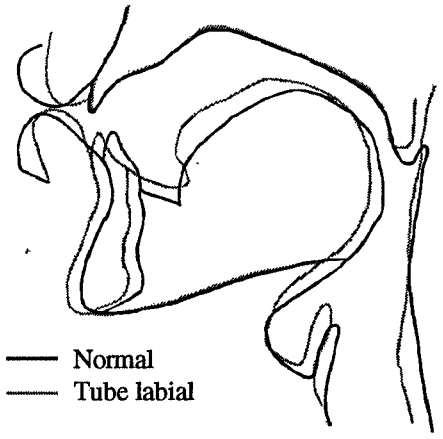


Figure 1 : Contours sagittaux référencés par rapport à l'incisive supérieure, en contexte normal et avec le tube labial de 2 mm.

• *Mesures acoustiques* : Les mesures effectuées sur le spectre du signal obtenu nous donnent pour les trois premiers formants : F1 = 271 Hz, F2 = 604 Hz et F3 = 1730 Hz. On peut aussi noter sur ce spectre une faible résonance aux alentours de 1200 Hz.

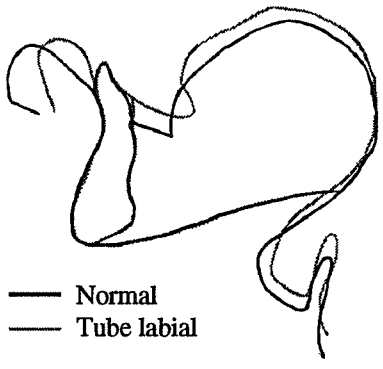


Figure 2 : Formes de la langue référencées par rapport à la mandibule.

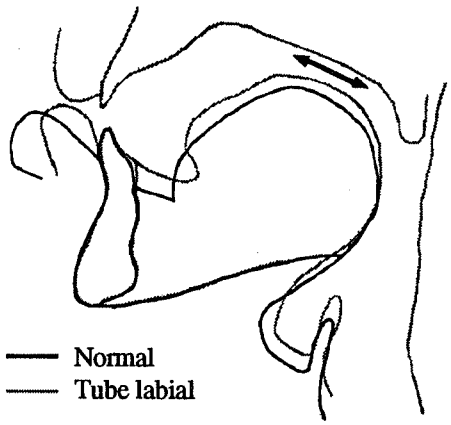


Figure 3 : Déplacements de la langue référencés par rapport à la mandibule dans la zone de constriction (repérée par la flèche).

• *Interprétations* : L'observation de la figure 2 permet de dissocier les mouvements respectifs de la mandibule et de la langue : l'avancée de la langue est essentiellement due à un déplacement de la mandibule, tandis que l'élévation résulte effectivement d'un mouvement de la langue relativement à la mandibule. Ce déplacement mandibulaire pourrait s'expliquer, tout comme la rétraction labiale, par un effort de préhension du petit tube labial, qui, compte tenu de sa petite taille et sous l'effet de la pression d'air au niveau des lèvres serait susceptible de glisser. Mais un tel mouvement ne parait cependant pas indispensable, une forte pression des lèvres semblant pouvoir être suffisante. On peut tout aussi bien l'imputer à une stratégie de compensation, qui, en induisant une avancée de la masse de la langue, réduit le volume de la cavité avant et contrecarre l'effet acoustique de la diminution de l'aire labiale. Mais ce mouvement, s'il était isolé, impliquerait un déplacement sensible de la constriction, et donc, par allongement de la cavité arrière, une diminution F3. C'est sans doute pour cette raison, que le locuteur a élevé sa langue dans sa partie dorsale (figure 2), ce qui a pour effet de rehausser la langue dans la zone vélaire pour recentrer le lieu de constriction dans cette zone ; c'est bien ce que tend à prouver la figure 3. Notons cependant, que si le lieu de constriction se situe effectivement toujours clairement dans la zone palato-vélaire, il est un peu plus antérieur que pour l'articulation du [u] normal : la compensation sur F3 a été insuffisante. Le couplage entre les cavités nasales et buccale, qui a pour effet acoustique de perturber la perception du [u] par l'apparition d'une fréquence aux alentours de 1200 Hz, ne semble pouvoir se justifier que par la nécessité de maintenir dans le conduit vocal un débit d'air suffisant, malgré la trop faible ouverture labiale.

Pour cette production sous contrainte, le maintien de la zone de constriction peut donc tout aussi bien s'expliquer par une stratégie de maintien de la cible acoustique (contrôle de F2 et F3) tout comme par une stratégie rigoureusement articulatoire (geste de préhension et contrôle de la zone de constriction). Notons cependant que la correction sur F2 est plus efficace que nécessaire et que celle du F3 est insuffisante, alors que ces deux corrections sont antagonistes ; on peut donc penser qu'une stratégie strictement acoustique aurait pu être plus efficace en minimisant la correction sur F2 pour améliorer celle du F3.

2.2.2. Cas du tube labial de 30 mm

• *Analyse de la géométrie du conduit vocal* : Le diamètre important du tube labial a pour première conséquence évidente, outre l'augmentation recherchée de l'aire labiale, une nette baisse de la mandibule (figure 4) ; on note dans le même temps un fort écrasement des lèvres sur le tube. Par ailleurs la forme de la langue se caractérise par une forte convexité dans la zone apicale, vraisemblablement imputable à un abaissement maximal de l'apex avec un maintien quasi parfait de la position et de l'aire de la constriction.

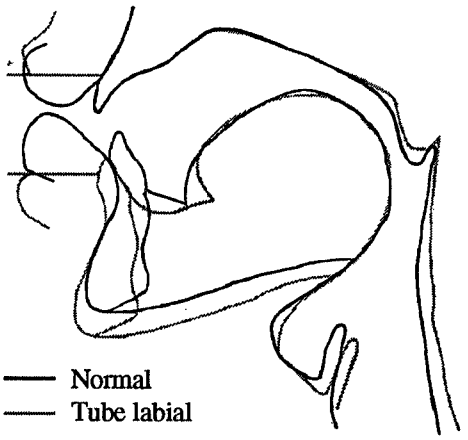


Figure 4 : Contours sagittaux référencés par rapport à l'incisive supérieure, en contexte normal et avec le tube labial de 30 mm.

• *Prédictions des conséquences acoustiques* : L'ouverture forcée des lèvres a pour conséquence acoustique majeure la disparition du Helmholtz formé par la cavité avant et les lèvres (mode de résonance initial de F2) ; cette cavité ne peut désormais plus résonner qu'en mode quart d'onde, ce qui autorise, en l'absence de modifications géométriques, une fréquence de résonance basse de l'ordre de 1000 Hz. L'effet acoustique de l'abaissement de l'apex n'est pas simple à expliquer ; cependant, dans le cadre d'une modélisation unidimensionnelle de la propagation des ondes, on peut supposer qu'en rendant plus abrupte le passage de la cavité à la constriction, cet abaissement aura tendance à augmenter la longueur acoustique équivalente de la cavité avant et à diminuer la fréquence affiliée à cette cavité ; il tendra donc à compenser l'effet de l'ouverture labiale. Le volume de la cavité arrière ainsi que la position de la constriction

n'ayant pas été modifiés, on ne devrait pas noter de changement significatif de F1 et de F3.

• *Mesures acoustiques* : Les résultats formantiques nous donnent les valeurs suivantes : F1 = 250 Hz, F2 = 920 Hz et F3 = 1958 Hz.

• *Interprétations* : Comme prévu, on observe la disparition de la fréquence basse autour de 700 Hz et l'apparition d'une fréquence légèrement inférieure à 1000 Hz. Cette dernière valeur confirme les prédictions faites ci-dessus sur les effets conjoints de l'ouverture labiale et de l'abaissement apical. La diminution (très nette pour F3) des deux autres formants n'est pas explicable au seul vu de nos données sagittales. F3 étant la résonance demi-onde de la cavité arrière, sa chute ne peut être attribuée qu'à un net allongement de cette cavité. La langue ne bougeant pas dans cette zone, on ne peut expliquer ce phénomène que par un abaissement du larynx, non observable sur nos tracés. Pour justifier cet abaissement, on peut reprendre les hypothèses émises par Wood (Wood, 1986) sur les corrélations entre l'aperture aux lèvres et la position du larynx : une forte ouverture des lèvres provoquerait une baisse significative du larynx. Dans ce cadre, il est vraisemblable que le tube labial a eu un effet secondaire lié à l'écartement important des lèvres : une baisse du larynx, et donc un allongement de la cavité arrière.

L'observation de la figure 5 permet de dissocier les mouvements respectifs de la mandibule et de la langue : la langue est très fortement relevée dans la partie dorsale. Il est très vraisemblable que l'abaissement de la mandibule est une conséquence secondaire de l'insertion du tube labial ; or une conséquence classique de l'abaissement mandibulaire consiste en un recul de la racine de la langue. C'est sans doute pour contrecarrer ce recul, que le locuteur a remonté de façon extrême sa langue afin de replacer la constriction dans la zone vélaire. Notons qu'acoustiquement cet effort ne semble pas devoir se justifier. En effet, compte tenu du nouveau mode de résonance quart d'onde de la cavité avant et de ses conséquences sur le patron formantique obtenu, un recul de la langue, induisant une constriction pharyngale et un allongement de la cavité avant, aurait été perceptivement bénéfique ; certes la cavité arrière en aurait alors été réduite, mais cela aurait pu aussi être compensé par une diminution de l'aire à la constriction, obtenue, dans cette zone pharyngale, précisément par un recul de la langue. Notons d'autre part que si, comme on peut le prévoir par l'analyse de F3, le larynx s'est abaissé, le recul de la langue n'aurait pas sensiblement réduit les dimensions de la cavité arrière.

Il semblerait donc que, pour cette production sous contrainte, le locuteur ait privilégié le maintien du lieu de constriction, au détriment de la précision acoustique. Certes l'abaissement apical va dans le sens d'une correction du formant associé à la cavité avant, mais il est difficile de savoir s'il résulte d'un contrôle volontaire ou s'il a pour effet de contribuer synergétiquement à la remontée importante de la langue dans sa partie dorsale.

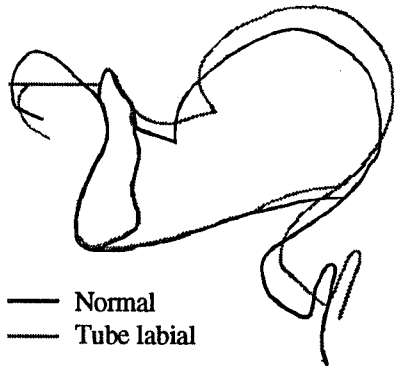


Figure 5 : Formes de la langue référencées par rapport à la mandibule.

3. CONCLUSION

Dans les deux conditions d'expérimentation – qu'il s'agisse d'une réduction ou d'une forte augmentation de l'aire labiale – la position et la taille de la constriction sont restées très sensiblement identiques à celles qui correspondent à une production naturelle. D'autre part, dans les deux cas, la réalisation acoustique n'est pas rigoureusement correcte, le troisième formant étant nettement trop faible ; cet abaissement est perceptivement sensible, si on se réfère aux données sur les seuils différentiels perceptifs (difference limens) publiés par Mermelstein (1978). Cependant on peut penser que cette modification perceptive, si elle altère la qualité de la voyelle, n'affecte pas sa classification phonétique en tant que [u] puisque le système phonologique français ne présente pas de concurrent direct du [u] sur la base d'une distinction sur F3 seul. Dans cette perspective, il devient nécessaire de faire une nette distinction entre nos deux cas expérimentaux : en effet avec le petit tube labial les valeurs de F1 et F2 sont tout à fait correctes, tandis que dans le cas du tube labial de gros diamètre elles ne sont en aucun cas admissibles pour la perception d'un [u] (cf. les seuils différentiels perceptifs). Ainsi donc les modifications de stratégies articulatoires observées pour le petit tube peuvent tout aussi rigoureusement être attribuées à un contrôle acoustique ou à un contrôle articulatoire. Il convient par conséquent de se pencher plutôt sur le cas du tube de gros diamètre.

Très clairement le maintien des caractéristiques du lieu de constriction s'est fait au dépens d'éventuelles corrections acoustiques. On aurait donc tendance à conclure que dans le contrôle de l'articulation, l'**espace articulatoire** constitue l'espace de contrôle distal, même si cela se fait au détriment de la précision acoustique. Le net écrasement des lèvres autour du tube de gros diamètre, attestant de la volonté manifeste de notre locuteur d'atteindre sa cible labiale, nous confirme dans cette conclusion. Cependant, nous l'avons souligné plus haut, l'abaissement de l'apex va dans le sens d'une compensation de l'effet acoustique de la perturbation

labiale. On peut donc fort bien envisager que le locuteur ait été sensible à l'altération perceptive et qu'il a tenté de développer une stratégie pour y remédier. Mais alors dans ce cas pourquoi n'a-t-il pas reculé très fortement sa langue ainsi que les modèles acoustiques conseilleraient de le faire ? Deux explications sont ici possibles : (1) il ne peut pas "casser" l'association que l'apprentissage de la parole a constituée entre voyelle et lieu de constriction, et il ne s'autorise de modifications qu'autour de cette position cible ; (2) compte tenu de la morphologie du conduit vocal et de la mobilité des articulateurs, la perturbation articulatoire est beaucoup trop importante pour être effectivement compensée, même par un recul maximal de la langue, et, face à cette incapacité, le locuteur se raccroche à la seule caractéristique encore possible, le lieu d'articulation.

Notre protocole expérimental, parce qu'il a ignoré l'observation de la phase d'apprentissage de l'**articulation** en présence du tube labial, ne nous permet pas de trancher sur ce point. Tout comme il ne nous permet pas de déterminer si l'abaissement apical est dû à un effet de synergie musculaire ou à un contrôle spécifique. Il est donc nécessaire de poursuivre l'expérience en intégrant un contrôle de l'apprentissage de l'**articulation** en présence des tubes labiaux ; nous le ferons sur plusieurs classes de locuteurs, naïfs et experts phonéticiens, en contrôlant l'information articulatoire ou acoustique accessible par le locuteur.

Notre expérience ne nous a donc pas réellement permis de trancher sur la primauté de l'espace acoustique ou de l'espace articulatoire dans le contrôle distal de l'articulation. Mais il nous permet d'ores et déjà d'affirmer que la zone de constriction constitue un **paramètre clef dans le contrôle** de la production des voyelles. Et cela représente un élément important, qu'il convient d'intégrer dans l'élaboration des modèles de contrôle de l'**articulation** et des processus de la perception de la parole.

REMERCIEMENTS

A Madame Martin du CHU de Grenoble (Service de Radiologie du Professeur Crouzet) pour le soin qu'elle a pris dans la réalisation des prises de vue radiologiques.

RÉFÉRENCES

BADIN P., PERRIER P., BOË L.J. & ABRY C. (1990), "Vocalic Nomograms : Acoustic and Articulatory Considerations upon Formant Convergences.", *J. Acoust. Soc. Am., 87,* 1290-1300.

BOË L.J., PERRIER P. & BAILLY G. (1992), "The Geometric Variables of the Vocal Tract Controlled for Vowel Production : Proposals for Constraining Acoustic-to-Articulatory Inversion.", *J. of Phonetics, 20,* 27-38.

CASTELLI E., PERRIER P. & BADIN P. (1989), "Caractérisation acoustique de la nasalité. A propos du premier formant nasal : quelques hypothèses résonn...ables.", *Bulletin du Laboratoire de la Communication Parlée* (No. 3, pp. 187-212). Grenoble : Institut National Polytechnique.

FANT G. (1960), *Acoustic Theory of Speech Production.* The Hague : Mouton.

FENG G., ABRY C. & GUERIN B. (1986), "The Nasopharyngeal Tract : A Target for Nasality. Acoustic Simulation versus Sweep-Tone Measurements.", *Actes du 12ème Congrès International d'Acoustique (A3-8).* Toronto.

GAY T., LINDBLOM B. & LUBKER J. (1981), "Production of Bite-block Vowels : Acoustic Equivalence by Selective Compensation.", *J. Acoust. Soc. Am., 69* , 802-810.

LALLOUACHE M. T. & WORLEY C. (1988), "Saisie, édition et traitement d'images de signaux articulatoires : lèvres et mâchoire.", *J. d'Acoustique, 1,* 215-220.

LALLOUACHE M.T. (1991), *Un poste "visage-parole" couleur ; Acquisition et traitement automatique des contours des lèvres.* Thèse de Doctorat non publiée, Institut National Polytechnique, Grenoble.

MAEDA S. (1979), "An Articulatory Model of the Tongue Based on a Statistical Analysis.", *J. Acoust. Soc. Am., 65,* S22.

MERMELSTEIN P. (1967), "Determination of the Vocal-Tract Shape from Measured Formant Frequencies.", *J. Acoust. Soc. Am., 41,* 1283-1294.

MERMELSTEIN P. (1978), "Difference Limens for Formant Frequencies of Steady-State and Consonant-Bound Vowels.", *J. Acoust. Soc. Am., 63* , 572-580.

SCHROEDER M. (1967), "Determination of the Geometry of the Vocal Tract by Acoustic Measurements.", *J. Acoust. Soc. Am., 41,* 1002-1010.

WOOD S. (1979), "A Radiographic Analysis of Constriction Locations for Vowels.", *J. of Phonetics, 7,* 25-43.

WOOD S. (1986), "The Acoustical Significance of Tongue, Lip and Larynx Maneuvers in Rounded Palatal Vowels.", *J. Acoust. Soc. Am., 80* , 391-400.

19es J.E.P. - Bruxelles, 19 au 22 mai 1992.

UNE NOUVELLE METHODE DE REDUCTION DES DONNEES ELECTROPALATOGRAPHIQUES

N. NGUYEN-TRONG & A. MARCHAL

Laboratoire Parole et Langage, CNRS, URA 261
Université de Provence, Aix-en-Provence, France

Résumé

Dans ce travail, nous présentons une méthode permettant de modéliser, à partir de données électropalatographiques, l'évolution des contacts entre la langue et le palais dans la production de la parole. Cette méthode consiste à représenter un palatogramme sous la forme d'une combinaison linéaire de *gestes élémentaires*, définis en soumettant les données EPG à une analyse en deux étapes (analyse spectrale suivie par une analyse factorielle). Nos résultats laissent penser qu'il serait possible de mettre à profit la méthode proposée, en vue d'élaborer un modèle tridimensionnel des mouvements de la langue dans la parole.

I. INTRODUCTION

On sait par de nombreux travaux que la manière dont la langue vient prendre appui contre le palais revêt une importance fondamentale dans l'émission d'une consonne comme /s/, /ʃ/ ou /l/ par exemple. Pour ces consonnes en effet, les mouvements articulatoires mis en oeuvre ne visent pas seulement à établir une constriction en un point du conduit vocal. La langue doit adopter une forme transversale concave (*grooving*) pour canaliser l'air expulsé par les poumons en direction des incisives inférieures (/s/), ou une forme convexe pour l'obliger à circuler le long de ses bords latéraux (/l/; voir Stone, 1991). Ces variations de forme dans le plan transversal marquent les limites d'une description articulatoire dans laquelle le conduit vocal se présente sous la simple forme d'une coupe sagittale. Elles montrent qu'il est indispensable de compléter cette description en étudiant avec précision la disposition spatiale des contacts qui s'établissent entre la langue et le palais.

Le problème abordé dans cet article est celui de modéliser l'évolution des appuis linguo-palatins dans la production de la parole, à partir de données obtenues par électropalatographie (EPG). On sait que l'EPG a déjà servi à de maintes reprises à définir des paramètres d'analyse articulatoire (ex.: position, largeur et longueur d'une constriction). Mais il n'existe encore (à notre connaissance du moins) aucun travail dont le but ait été de "synthétiser" des palatogrammes, par l'intermédiaire d'un modèle statistique analogue à ceux que l'on a mis au point pour reconstituer les mouvements articulatoires dans le plan sagittal (Harshman *et al.*, 1977, Maeda, 1990). Notre propre travail est donc destiné à introduire ce type de modèle dans le domaine de l'électropalatographie. Il vise à définir, par une investigation statistique, un petit nombre de paramètres qui permettent de contrôler la configuration des zones de contact langue-palais, et qui puissent être intégrés dans une étape ultérieure à un modèle *tridimensionnel* des mouvements accomplis par la langue dans la parole.

II. PLAN EXPERIMENTAL

Rappelons brièvement que l'EPG (Marchal, 1988) repose sur l'utilisation d'un palais artificiel en résine porté par le sujet, et sur la surface duquel se trouvent réparties plusieurs dizaines d'électrodes. Lorsque la langue entre en contact avec une électrode, celle-ci fonctionne comme un commutateur laissant passer un courant de très faible voltage que l'on fait recevoir au sujet. L'image montrant l'emplacement des électrodes en position «on» et en position «off» sur le palais artificiel, est appelée un palatogramme. Le système utilisé dans le présent travail a été mis au point à l'Université de Reading (Hardcastle *et al.*, 1989). Dans ce système, le palais artificiel comporte 62 électrodes, et la fréquence d'échantillonnage du signal EPG est de 200 Hz.

Le corpus se compose de 149 items qui se répartissent en quatre catégories. La première catégorie est formée par les voyelles isolées /i/, /a/, /u/, /ɛ̃/, /ɑ̃/. Les trois autres catégories ont été constituées en combinant deux voyelles cardinales (/i/, /a/ ou /u/) avec une consonne simple tirée de l'ensemble /p, b, t, d, k, g, s, z, ʃ, ʒ, n, l/ (séquences

VCV), ou bien un groupe de deux consonnes: /tʃ/, /st/ ou /kl/ (séquences VCCV), ou bien encore un groupe de trois consonnes: /skl/ (séquence VCCCV). Les données EPG sur lesquelles portent nos analyses sont relatives à un locuteur, de sexe féminin, et elles ont été extraites de la base multi-paramétrique EUR-ACCOR (Marchal *et al.*, 1991).

Dans une phase préliminaire, une opération de tri a été mise en oeuvre qui était destinée à éliminer les palatogrammes apportant peu d'information. Nous avons considéré qu'un palatogramme devait pour être conservé répondre à deux critères: *a)* présenter un nombre de contacts supérieur ou égal à 6; *b)* présenter une différence portant sur au moins deux électrodes avec le palatogramme qui le précédait immédiatement sur l'axe temporel. Ce second critère est motivé par le fait que le signal EPG varie de manière relativement lente dans la production de la parole, et que les appuis langue-palais peuvent conserver la même configuration pendant un laps de temps assez long (quelques dizaines de ms). Le nombre total de palatogrammes sélectionnés est de 2774.

III. PREMIERE DECOMPOSITION EN IMAGES PRIMITIVES

La méthode que nous nous proposons d'adopter pour extraire les paramètres contrôlant l'évolution des appuis langue-palais, se compare dans sa première phase à une transformée de Fourier discrète. On sait bien sûr qu'une telle transformée est utilisée pour décomposer un signal échantillonné unidimensionnel (signal acoustique par exemple) en une somme de composantes sinusoidales dont les fréquences respectives sont des multiples entiers d'une fréquence de base. On sait aussi qu'il est possible d'étendre le domaine d'application des transformées de ce type à un signal à deux dimensions, c'est-à-dire à une image, discrétisée sous la forme d'une grille de points. Dans ce second cas, les composantes dont le signal se présente comme une somme pondérée ne sont plus constituées par des sinusoides, mais par des images de base, ou images primitives (Clarke, 1985). Il est clair qu'un palatogramme est assimilable à une image (composée de points «noirs» ou «blancs» selon que la langue est en contact ou non avec les électrodes correspondantes), et qu'il est donc susceptible d'être soumis à une transformée bidimensionnelle de type Fourier. L'intérêt de cette opération réside dans le fait que les images primitives peuvent donner lieu, sous certaines conditions, à une interprétation en termes articulatoires, comme cela est montré plus loin.

La transformée que nous avons choisi d'utiliser est la transformée en cosinus discrète (*discrete cosine transform* ou DCT), qui présente sur la transformée de Fourier proprement dite l'avantage d'être plus facile à mettre en oeuvre car elle ne fait intervenir que des nombres réels. La DCT, appliquée à une image comportant un même nombre N de lignes et de colonnes se calcule par la formule suivante:

$$C_{pq} = C_o \, 2/N \sum_{k=0}^{N-1} \sum_{l=0}^{N-1} f_{kl} \, t_{ql} \, t_{pk}$$

avec:

$p = 0,\ldots,N-1;\ q = 0,\ldots,N-1$

$t_{ql} = \cos(2l+1)q\,\pi/2N$

$t_{pk} = \cos(2k+1)p\,\pi/2N$

$C_o = 1/\sqrt{2}$ si $p=q=0$, et 1 sinon

On désigne par f_{kl} la valeur de l'élément situé au point d'intersection de la ligne k et de la colonne l dans l'image, et par C_{pq} celle du coefficient appartenant à la ligne p et à la colonne q dans le tableau où se rangent les résultats de la transformée.

Ce tableau comporte lui-même N lignes et N colonnes. On peut interpréter les coefficients dont il se compose comme des valeurs d'énergie relatives à NxN composantes spectrales réparties à intervalles égaux sur une échelle de fréquences spatiales. Les composantes basses fréquences sont à mettre en relation avec les variations de couleur (ou de niveau de gris) étalées sur de vastes zones de l'image; les composantes hautes fréquences correspondent à des variations plus abruptes, qui se produisent en un point bien déterminé dans l'espace (bords ou frontières). La composante continue C_{00} (*DC component*) représente l'énergie moyenne de l'image.

Une fois que les coefficients C_{pq} ont été calculés, il est possible de reconstituer l'image initiale par une DCT inverse dont la formule est la suivante:

$$f_{kl} = C_o \, 2/N \sum_{p=0}^{N-1} \sum_{q=0}^{N-1} C_{pq} \, t_{ql} \, t_{pk}$$

avec: $k = 0,\ldots,N-1;\ l = 0,\ldots,N-1$

En fait, on peut ne pas faire entrer en jeu dans ce calcul les composantes spectrales de plus haute fréquence sans altérer notablement la qualité de la reconstitution, dès lors que l'image ne présente pas de discontinuités trop marquées. Cette propriété fait d'une transformée de type DCT un outil de compression de l'information (de réduction des données) pouvant servir à recoder une image sous la forme d'un petit ensemble de coefficients.

Nous avons ainsi cherché en premier lieu à déterminer le nombre de composantes spectrales nécessaires pour reconstituer un palatogramme de manière satisfaisante. Les images EPG décrites dans la section II ont donc été soumises chacune à une DCT, puis reconstituées grâce à une DCT inverse faisant appel à des composantes dont nous avons fait varier le nombre entre 15 et 30, en prenant celles qui appartenaient à la zone des fréquences les plus

basses. Précisons que les palatogrammes ont été considérés comme des images 8x8, bien que le palais artificiel ne soit muni que de 62 électrodes (les électrodes disposées sur le bord antérieur du palais sont au nombre de 6). Les 2 éléments de l'image correspondant à des électrodes inexistantes ont systématiquement reçu pour valeur 0.

Les résultats sont représentés sur la figure 1 par des courbes illustrant la manière dont les données se distribuent selon le nombre d'erreurs relevées dans l'image reconstituée, c'est-à-dire le nombre d'éléments possédant une valeur opposée à leur valeur initiale (0 au lieu de 1 ou 1 au lieu de 0).

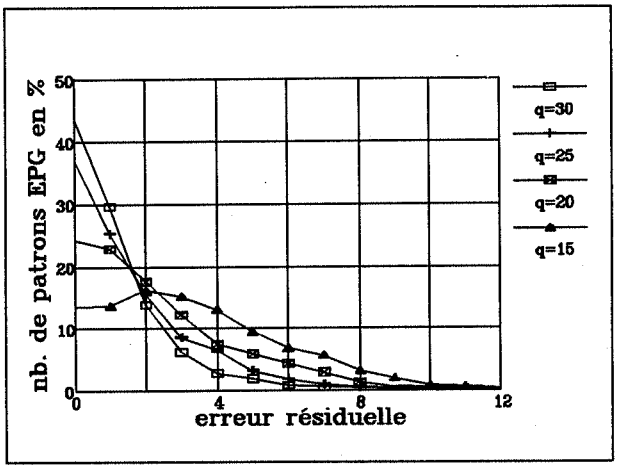


FIGURE 1: Distribution des données EPG (n=2774) selon le nombre d'erreurs commises en reconstituant la configuration des appuis langue-palais par une DCT inverse. La reconstitution s'est fondée sur un nombre de composantes spectrales allant de q=15 à q=30.

On constate ainsi que le pourcentage de palatogrammes pour lesquels le nombre d'erreurs est supérieur à 5 (sur 64 éléments) se montre extrêmement réduit lorsque la reconstitution fait intervenir les 25 ou les 30 premières composantes. En fait, la qualité de la reconstitution ne semble véritablement se dégrader qu'à partir du moment où le nombre de composantes utilisées est ramené à 15. De tels résultats apportent une confirmation statistique à l'idée selon laquelle les images palatographiques présentent une forte redondance, tenant aux corrélations qui s'établissent en elles entre deux éléments voisins. La langue ne peut à l'évidence établir de contact avec chaque électrode indépendamment des autres. Nos analyses font apparaître que les palatogrammes sont ainsi susceptibles de donner lieu à une compression d'information relativement importante. Mais on doit admettre que le nombre de composantes spectrales assurant une bonne reconstitution de ces palatogrammes, reste assez élevé (entre 20 et 25). Selon nous, ce phénomène s'explique en partie par le fait que les données EPG sont codées de manière binaire (1: contact, 0: absence de contact). Par suite, les appuis langue-palais dessinent sur chaque image une configuration dont les contours extrêmement marqués peuvent conférer aux composantes spectrales de haute fréquence une importance non négligeable. Il n'est pas exclu de penser qu'un lissage préalable des données permettrait à la DCT de produire de meilleurs résultats.

Nous avons indiqué plus haut que ce type de transformée amenait à représenter une image sous la forme d'une somme pondérée d'images «primitives». Cela signifie qu'il suffit simplement pour reconstituer cette image d'additionner les primitives après avoir multiplié leurs éléments respectifs par un coefficient de pondération (dont la transformée directe vise précisément à établir la valeur). La figure 2 représente les 16 premières des 64 primitives que la DCT fait correspondre à une image 8x8. Le blanc est ici associé aux valeurs comprises entre -1 et 0, le noir aux valeurs comprises entre 0 et 1.

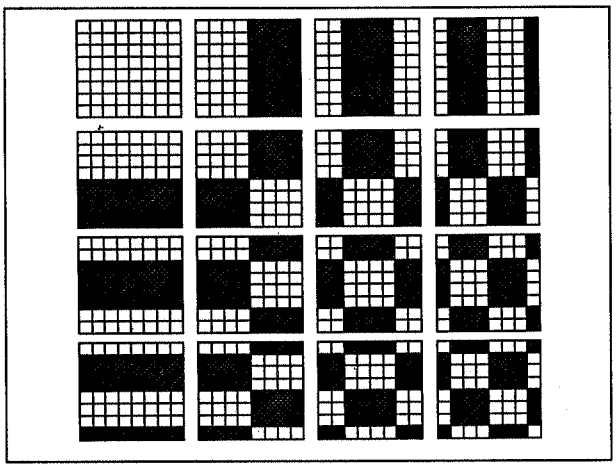


FIGURE 2: 16 premières des 64 primitives associées à une image 8x8 dans la transformée en cosinus discrète (d'après Clarke, 1985).

Il est facile de constater que ces primitives se prêtent, pour certaines d'entre elles, à une interprétation en termes articulatoires. La primitive située sur la ligne 1, colonne 1 reflète l'énergie moyenne de l'image; elle peut donc fournir une indication directe sur le nombre total de contacts établis entre la langue et le palais. La primitive se trouvant sur la ligne 1, colonne 2 instaure une opposition entre la partie droite et la partie gauche de l'image; elle est ainsi utilisable pour modéliser une asymétrie latérale dans les appuis linguo-palatins. La primitive située sur la ligne 1, colonne 3 représente une suite de deux passages par zéro sur l'axe horizontal; elle permet ainsi de différencier les palatogrammes selon que la langue vient s'appuyer contre les bords latéraux du palais, ou bien dans sa partie centrale. Enfin, la primitive située sur la ligne 2, colonne 1 représente un passage par zéro dans le sens vertical; elle peut donc servir à déterminer l'emplacement des contacts

langue-palais sur l'axe antéro-postérieur.

Il est en fait possible d'interpréter ces primitives comme des traits articulatoires: haut/bas, droite/gauche, latéral/central, avant/arrière, et les coefficients dont elles sont munies comme des indicateurs signalant si un trait se trouve ou non présent dans l'image analysée. L'un des avantages de la DCT est que de tels indicateurs peuvent tous être calculés à partir d'une même formule de base.

III. SECONDE DECOMPOSITION EN GESTES ELEMENTAIRES

La figure 2 montre cependant que de nombreuses primitives, lorsqu'elles sont prises isolément, ne rappelent aucune configuration EPG réalisable. Sous un certain aspect en fait, ces images nous semblent comparables aux éléments d'une sorte de jeu de construction, qu'il est nécessaire de combiner les uns avec les autres pour aboutir à des formes dotées d'une signification phonétique. Le problème qui se pose est que le nombre de combinaisons possibles est extrêmement élevé. En outre, il serait difficile d'explorer ces combinaisons sans un critère permettant de reconnaître celles qui s'accordent le mieux avec les patterns EPG observés, c'est-à-dire celles qui possèdent le pouvoir descriptif le plus fort.

La solution que nous avons adoptée consiste simplement à soumettre les 25 premiers coefficients fournis par la DCT pour chaque image électropalatographique incluse dans notre matériel, à une analyse factorielle (analyse en composantes principales). Cette analyse est destinée à mettre en évidence de possibles corrélations entre les différents coefficients. Elle doit en d'autres termes nous indiquer si les primitives associées à ces coefficients varient d'une manière conjointe, d'un palatogramme à l'autre. Dans l'affirmative, on pourra considérer que ces primitives se combinent dans les palatogrammes sous le contrôle d'un ensemble plus réduit de paramètres, qui ne sont pas autre chose que des facteurs, au sens que ce terme revêt dans une analyse factorielle.

Les résultats de l'analyse sont résumés sur la figure 3, sous la forme d'une courbe illustrant l'accroissement du pourcentage cumulé de variance expliquée entre le premier facteur (c'est-à-dire celui qui absorbe la part la plus importante de la variance présentée par les données) et le dernier. On constate que ce pourcentage augmente de façon assez lente, et qu'une douzaine de facteurs sont nécessaires pour expliquer environ 80% de la variance initiale. La figure 3 donne ainsi à penser que les primitives ne «s'imbriquent» pas les unes dans les autres aussi étroitement qu'on pourrait le supposer. Mais il est également possible que notre méthode d'analyse doive être améliorée en plusieurs points, qui seront examinés dans la section suivante. Signalons que la part de variance expliquée est de 17.6% pour le premier facteur, et de 11.2% pour le deuxième facteur.

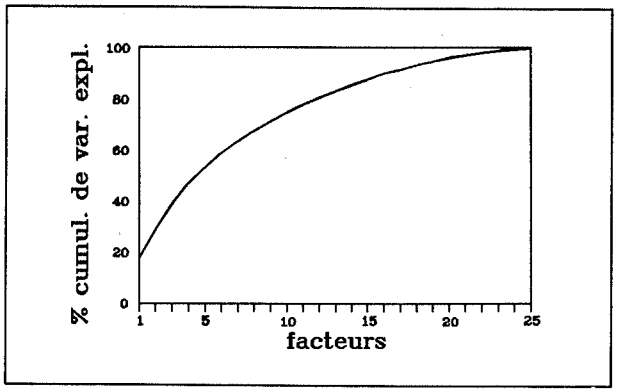


FIGURE 3: croissance du pourcentage cumulé de variance expliquée entre le premier et le dernier facteur.

Ces deux premiers facteurs s'interprètent facilement grâce à une procédure dont les résultats sont présentés sur la figure 4, et qui offre la possibilité de déterminer la manière dont les appuis linguo-palatins se modifient lorsque l'on fait varier chaque facteur entre -2 et +2 écarts types. Cette procédure a consisté à calculer, pour chaque valeur attribuée à un facteur, 25 coefficients spectraux, par une opération qui peut être définie comme l'inverse d'une analyse factorielle (voir par exemple Harshman *et al.*, 1977); les coefficients ont alors été à leur tour transformés en une image EPG, par l'intermédiaire d'une DCT inverse. Il est important de souligner que les mouvements articulatoires illustrés sur la figure n'ont pas été directement observés. Ils constituent ce qu'il est possible d'appeler un *geste élémentaire*, pour reprendre une expression proposée par Maeda (1990). Ce geste a été isolé grâce à une analyse appliquée en deux étapes (DCT + analyse factorielle) aux données EPG dont se compose notre matériel. La procédure de «synthèse» décrite ci-dessus nous permet de voir les gestes se réaliser in abstracto en quelque sorte.

Le geste associé au premier facteur s'apparente à un mouvement assez ample d'abaissement de la langue. Dans ce mouvement, la langue commence par relâcher son appui au centre du palais, pour se détacher ensuite de ses bords latéraux. On constate également que le nombre de contacts diminue d'abord à la hauteur des alvéoles, et en un second temps dans la partie postérieure. Les palatogrammes laissent apparaître une intéressante asymétrie latérale qu'un examen des données nous a permis d'identifier comme étant une caractéristique individuelle. Le geste associé au deuxième facteur se présente comme un mouvement de l'apex visant à établir un barrage semi-circulaire sur les bords du palais. On voit très nettement une occlusion se former au niveau antérieur selon un mécanisme analogue à la fermeture d'une valve

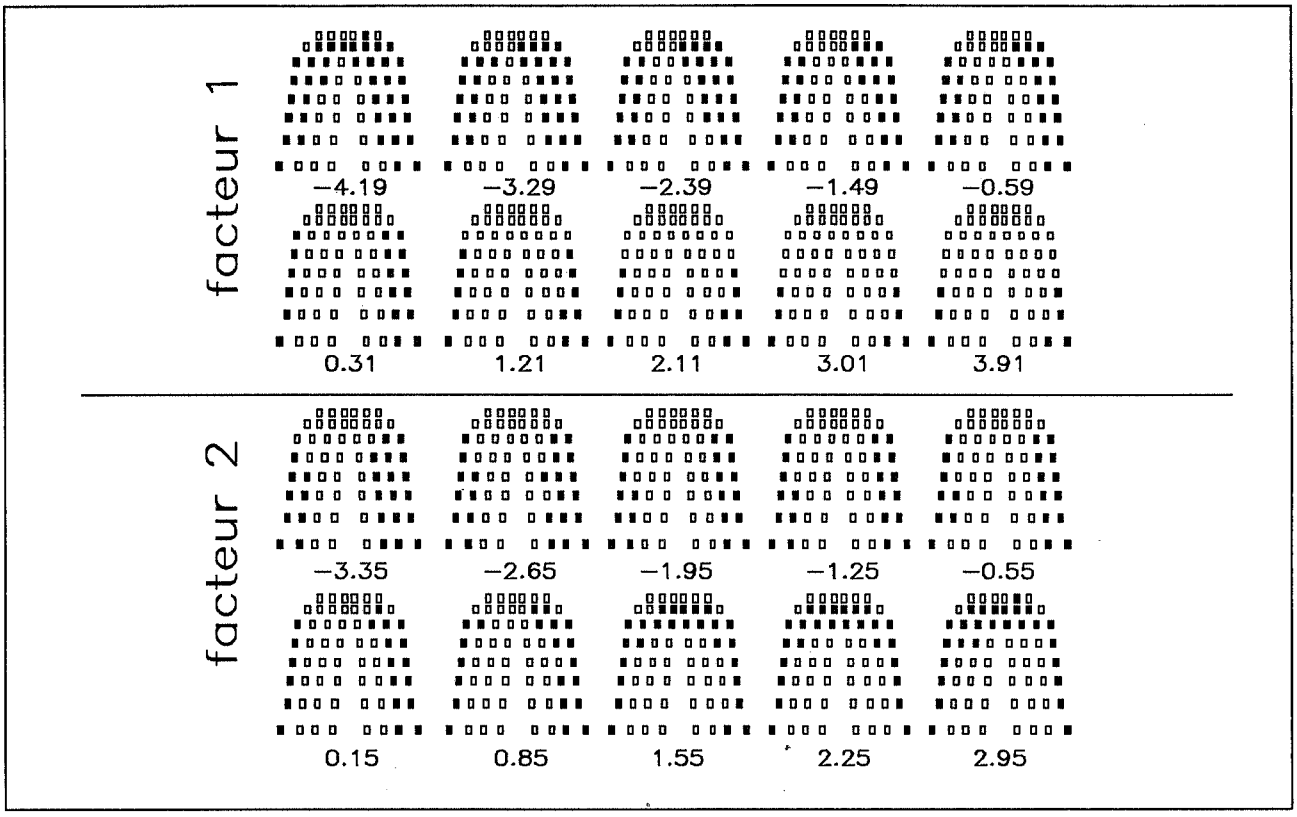


FIGURE 4: modifications théoriques des appuis linguo-palatins obtenus en faisant varier le facteur 1 entre -2 et +2 écarts types. La valeur du facteur pour chaque image EPG se trouve indiquée au-dessous de cette image.

(progression des appuis de la périphérie vers le centre; cf. Keller, 1991). La langue prend d'abord appui contre les bords latéraux avant de venir adhérer à la voûte palatine. On note comme pour le premier facteur une certaine asymétrie dans la disposition des contacts de part et d'autre de la ligne médiane.

IV. RECONSTITUTION DES IMAGES EPG INITIALES

Le modèle (au sens statistique du terme) que nous venons de décrire a été soumis à un premier test, qui a consisté à reconstituer les images EPG initiales (n=2774) à partir des 7 premiers facteurs. La figure 5 fait apparaître que les résultats obtenus sont assez encourageants. Le nombre d'erreurs produites par le modèle reste inférieur à 10 environ pour la majeure partie des données. En outre, il semble que les erreurs n'aient pas eu lieu au hasard, et que les palatogrammes les plus difficiles à reconstituer puissent être reconnus à certaines caractéristiques bien définies. Nous avons constaté en particulier que les patterns EPG correspondant à des mouvements rapides de la langue (transitions) donnaient souvent lieu à un taux d'erreur élevé. Ce phénomène n'a rien d'étonnant puisque les patterns dont le le rôle a été le plus important dans le calcul des facteurs, sont évidemment ceux qui se répètent le plus grand nombre de fois dans notre matériel, c'est-à-dire ceux qui correspondent à des phases de stabilité articulatoire.

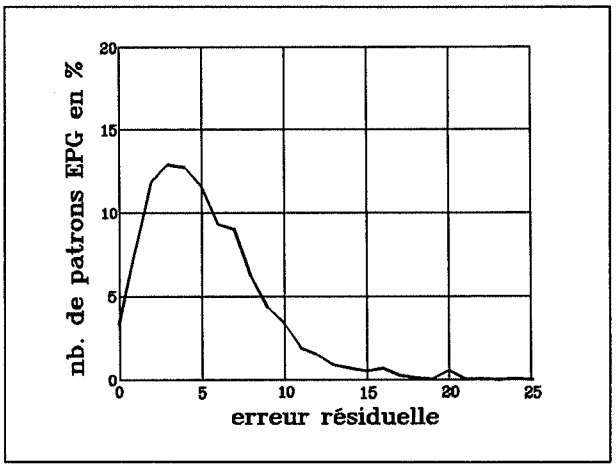


FIGURE 5: distribution des données EPG (n=2774) selon le nombre d'erreurs commises en reconstituant la configuration des appuis langue-palais à partir des 7 premiers facteurs.

V. DISCUSSION

Nous avons indiqué qu'il n'a pas été possible d'utiliser un nombre de coefficients DCT inférieur à 20 dans le modèle, sans provoquer une augmentation notable des erreurs dans la reconstitution des images EPG. On peut penser que ce problème revêtirait une importance moins grande si les données étaient soumises à un lissage préliminaire, selon une méthode du type moyenne glissante par exemple. En calculant, pour chaque image EPG acquise à un instant *t*, la valeur moyenne de chaque contact dans un intervalle centré en *t* (et dont la taille reste à déterminer), on parviendrait à "adoucir" les contours des appuis langue-palais sur l'image, et donc à faire chuter l'énergie présentée par le spectre correspondant en hautes fréquences. Par ailleurs, il est apparu que le nombre de facteurs requis pour expliquer la majeure partie de la variance parmi les coefficients DCT, était relativement élevé (entre 12 et 15). Cet autre problème est en partie attribuable au fait que le corpus comportait des consonnes et des groupes de consonnes extrêmement variés. Il est vraisemblable qu'une analyse portant sur un corpus plus homogène aurait abouti à des résultats plus satisfaisants. Remarquons en outre que les indices que nous avons utilisés afin de tester notre modèle sont des indices numériques (pourcentage de variance expliquée en ce qui concerne les facteurs, taux d'erreur dans la reconstitution des images pour les coefficients spectraux), dont la portée reste limitée. On peut en effet fort bien imaginer un modèle donnant lieu à un taux d'erreur important tout en préservant les principales caractéristiques *qualitatives* des palatogrammes à reconstituer. Il est clair que nos critères d'évaluation ont vraisemblablement à être améliorés sur ce point.

En résumé, ce travail laisse penser qu'il est possible de reconstituer l'évolution des appuis entre la langue et le palais dans la production de la parole, au moyen d'un modèle à deux niveaux. Le premier niveau est celui des images EPG primitives, dont chaque palatogramme peut se représenter comme une somme pondérée, grâce à une transformée en cosinus discrète. Le second niveau est celui des facteurs EPG reflétant les relations d'interdépendance qui s'établissent entre images primitives dans les données analysées. Le terme de geste élémentaire est employé ici pour désigner la façon dont les appuis linguo-palatins se modifient sous le contrôle d'un facteur déterminé, par l'intermédiaire des primitives.

On peut juger que la DCT et l'analyse factorielle sont utilisées l'une à la suite de l'autre dans un même but qui est celui de montrer que les configurations EPG présentent un nombre limité de degrés de liberté. Cependant, il existe entre ces deux analyses une différence essentielle tenant au fait que la DCT est appliquée localement (c'est-à-dire à chaque palatogramme pris séparément), tandis que l'analyse factorielle s'étend à un ensemble de données, dont elle permet de connaître les dimensions de variance maximale dans l'espace des paramètres descriptifs choisis.

Par ailleurs, il convient de bien souligner que le matériel servant de base empirique à notre modèle est exclusivement formé par des données EPG. Cela signifie que les paramètres contrôlant la disposition des contacts entre la langue et le palais sont déterminés par inférence, à partir de ces données. De tels paramètres peuvent ainsi être considérés comme se trouvant dissimulés dans une sorte de boîte noire, que nous cherchons à étudier de l'extérieur, à travers les mouvements articulatoires dont elle commande la mise en oeuvre. Il est selon nous indispensable d'associer cette investigation de type "bottom-up" à une investigation en sens inverse, de type "top-down", visant à déterminer des paramètres de contrôle moteurs à partir de données anatomiques, physiologiques et neuro-physiologiques. Le modèle de Stone (1991) par exemple, dans lequel la langue se subdivise en blocs fonctionnels semi-indépendants, demande à être corroborré par une analyse montrant l'existence sur un palatogramme de solidarités fonctionnelles, c'est-à-dire de zones de corrélations entre contacts. C'est précisément pour mettre en évidence de telles zones que la méthode présentée dans ce travail a été mise au point.

VI. REMERCIEMENTS

Ce travail a été entrepris dans le cadre du projet ESPRIT / ACCOR.

VII. BIBLIOGRAPHIE

Clarke, R.J. (1985). *Transform Coding of Images* (Academic Press, London).

Hardcastle, W., Jones, W., Knight, C., Trudgeon, A., et Calder, G. (1989). «New developments in electropalatography: a state-of-the-art report,» *Clinical Linguistics & Phonetics* 3, 1-38.

Harshman, R., Ladefoged, P., et Goldstein, L. (1977). «Factor analysis of tongue shapes,» *J. Acoust. Soc. Am.* 62, 693-707.

Keller, E., et Gabioud, B. (1991). «Testing a simplified articulatory model for speech timing predictions,» conférence présentée au *Symposium ACCOR sur la coarticulation*, Venise, 24-26 octobre 1991.

Maeda, S. (1990). «Compensatory articulation during speech: evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model,» *in Speech Production and Speech Modelling*, édité par W.J. Hardcastle et A. Marchal (Kluwer, Dordrecht), pp.131-149.

Marchal, A. (1988). *La Palatographie* (CNRS, Paris).

Marchal, A., *et al.* (1991). «EUR-ACCOR: the design of a multichannel database,» *in Actes du XIIème Cong. Int. des Sc. Phonétiques*, Aix-en-Provence, 19-24 août 1991, vol.5, pp.422-425 [article présenté au nom du consortium ACCOR].

Nguyen, N., et Marchal, A. (1991). «A note on EPG data reduction,» *ACCOR 2nd Periodic Progress Report*, vol.3.

Stone, M. (1991). «Toward a model of three-dimensional tongue movement,» *J. Phonetics* 19.

19es J.E.P. - Bruxelles, 19 au 22 mai 1992.

TRAITEMENT LINGUISTIQUE ET PHONETIQUE DU FRANÇAIS DANS UN SYSTEME DE SYNTHESE DE LA PAROLE MULTILINGUE

Luc Mortier, Bert Van Coile[1]

Lernout & Hauspie Speech Products, Rozendaalstraat 14, B-8900 Ieper, Belgique

Résumé

Cet article est consacré au développement de systèmes de synthèse de parole à partir d'un texte quelconque. Nous indiquerons divers problèmes linguistiques relatés: la conversion de graphèmes en phonèmes, l'analyse lexicale et l'analyse syntaxique. La méthode adoptée chez *Lernout & Hauspie Speech Products* (LHS) est présentée et illustrée par des exemples tirés de notre système de synthèse de la parole français. Nous esquisserons aussi notre environnement de développement linguistique, qui est indispensable à la réalisation systématique et rapide de notre stratégie.

INTRODUCTION

La conversion de texte en parole par ordinateur est une tâche qui demande la mise en oeuvre de toute une gamme de connaissances linguistiques et technologiques. Vu sous un angle commercial, un tel système doit pouvoir produire de la parole synthétique de qualité aussi haute que possible, tout en respectant les limites imposées par le hardware utilisé. Pour notre système de base, ce hardware est relativement restreint (processeurs: 80186 et TMS-C25; mémoire: 750 Ko). Dès lors, un équilibre doit être trouvé entre ces restrictions (imposées par le marché et par la technologie) et les besoins linguistiques et acoustiques.

Beaucoup d'applications, spécialement en Europe, exigent un produit multilingue. Chez LHS, nous développons actuellement un synthétiseur de parole pour sept langues. Par conséquent, une stratégie efficace et uniforme s'impose.

Dans la plupart des systèmes existants, la transformation de texte en parole est réalisée à travers les étapes suivantes:

- Normalisation du texte
- Conversion de lettres (graphèmes) en phonèmes
- Analyse lexicale, syntaxique (et sémantique). Cette analyse est nécessaire pour assigner une prononciation correcte aux mots hétérophones (par exemple: président, portions) et pour obtenir une bonne synthèse prosodique.
- Synthèse segmentale: détermination des caractéristiques spectrales. Les systèmes développés à LHS sont basés sur la concaténation de diphones et triphones.

[1]également attaché à l'Université de Gand (Laboratoire d'Electronique et de Métrologie).

- Synthèse prosodique: insertion de pauses, calcul de la durée des phonèmes et du contour intonatif. Ces aspects sont réalisés par application de règles.
- Synthèse proprement dite du signal acoustique.

Dans ce qui suit, nous traiterons surtout la conversion de lettres en phonèmes, et l'analyse lexicale et syntaxique.

CONVERSION DE MOTS ISOLES EN PHONEMES

La conversion de mots en phonèmes a déjà été abordée de plusieurs manières :

- par dictionnaires de mots fléchis
- par règles de prononciation (créées manuellement [2,4] ou par apprentissage automatique [11,5])
- par décomposition morphologique (règles plus dictionnaire de morphèmes) [1,6]
- par réseaux neuraux [8]

En général, l'utilisation de fichiers lexicaux phonétiques offre certains avantages: vitesse d'exécution, exactitude de la transcription phonétique et simplicité de la méthode. En outre, ces lexiques peuvent être enrichis de connaissances linguistiques, telles la catégorie grammaticale. Ils présentent également des inconvénients importants. Ces fichiers occupent beaucoup de mémoire et sont toujours incomplets. En plus, cette approche est statique et donc incapable de faire face au dynamisme de la langue (néologismes, mots composés). Une solution à ces problèmes peut être l'approche morphologique. Le lexique y est réduit aux morphèmes de base (lexicaux et grammaticaux). Une stratégie, basée sur la consultation intensive du lexique et sur des règles, est utilisée pour la décomposition des mots en morphèmes. Finalement, la transcription phonétique du mot est réalisée à base des transcriptions des morphèmes constitutifs. Or cette approche exige une certaine puissance de calcul et une quantité de mémoire non négligeable. Par exemple, le système anglais américain décrit dans [6] utilise un fichier de 43000 morphèmes; programme et données représentent 900 Ko de mémoire.

Quant à la conversion par règles de prononciation, il est clair qu'elle aboutit à une transcription phonétique pour chaque mot. Cependant des erreurs de prononciation sont inévitables. Remarquez que cette approche-ci et la méthode des lexiques à mots fléchis présentent des caractéristiques contraires. Les résultats qu'on peut obtenir grâce aux règles de prononciation, varient d'une langue à l'autre. Le degré de réussite est évidemment très élevé pour des langues comme l'espagnol et le coréen, qui sont caractérisées par un rapport orthographe-phonétique très régulier. Nous avons également obtenu de hautes performances pour le néerlandais, le français et l'allemand. Pour chacune de ces langues, quelques 300 règles suffisent pour obtenir moins de 10 % de mots incorrects (tenant compte des erreurs de phonétisation et d'accentuation) sur les 10000 mots fléchis les plus fréquents. Mais nous n'avons pas pu atteindre ce chiffre pour l'anglais, même en utilisant plus de règles.

Il est clair que la seule utilisation de règles de prononciation ne suffit pas à résoudre le problème de conversion phonétique globalement. C'est pourquoi nous avons adopté une stratégie hybride, utilisant conjointement les éléments suivants:

- un convertisseur par règles
- un dictionnaire de mots à classe fermée
- un dictionnaire de mots fléchis à haute fréquence
- analyse morphologique, par règles graphotactiques et/ou par lexique morphologique plus règles
- un dictionnaire d'exceptions

Le fait d'avoir différentes stratégies à notre disposition nous permet de doser celles-ci en fonction des problèmes particuliers de chaque langue. Par exemple: la décomposition morphologique (non-exhaustive) de l'anglais a été choisie en raison de l'irrégularité orthographi-

que de cette langue, et donc pour augmenter la performance de la conversion phonétique, aussi bien que la vitesse d'exécution. En même temps, cette approche permet d'identifier la nature grammaticale de tous les mots dérivables des racines incluses (verbes, adjectifs et noms dérivés, homographies).

ANALYSE LEXICALE ET MORPHOLOGIQUE

Dans notre système français de synthèse de parole, l'analyse morphologique sert à identifier la catégorie grammaticale des mots. Cette information est e.a. indispensable à l'analyse syntaxique et à la synthèse prosodique de la phrase.

Le système utilise deux listes de *morphèmes* (La notion de morphème n'est pas utilisée au sens strictement linguistique.). Dans la version actuelle, aucune transcription phonétique n'est incorporée dans ces listes de morphèmes, puisque les règles de conversion sont suffisamment performantes. Voici un extrait des listes morphologiques (à gauche: racines, à droite terminaisons).

arm	ai
port	ais
fin	e
meur	es
mour	ions
pren	is
prend	issent
recev	rez
reçoi	t
reç	ûmes
...	...
au total:	
4760 +	101

Le fichier réel contient aussi des codes morphologiques. L'algorithme consiste à rechercher d'abord une terminaison possible et de vérifier ensuite si le reste du mot est une racine. Des règles de compatibilité entre type de racine et terminaison assurent que le découpage est exact. Par exemple, ces règles acceptent la décomposition *fin+is* et rejettent *fin+es*, puisque *es* est incompatible avec cette racine verbale. Certaines racines ont un code grammatical ambigu. Des règles aident là aussi à l'élimination ou la réduction de l'ambiguïté. Comparez:

arm+ée	nom fém. sg.	part. passé
arm+es	nom fém. pl.	verbe
arm+erons	verbe	

Les ambiguïtés restantes doivent être traitées au niveau syntaxique.

Remarquons encore que nous employons aussi d'autres moyens pour déterminer la classe grammaticale des mots:

- la consultation de lexiques à mots fléchis:
  - mots de haute fréquence (360 entrées)
  - mots à classe fermée (328 entrées). Ce lexique contient également des groupes de mots comme *en dépit de* et *jusqu'à ce que.*
  - mots hétérophones (46)
- la détection de suffixes prédictifs

ANALYSE SYNTAXIQUE

Après l'analyse lexicale et la conversion de mots en phonèmes, notre système fait appel à des règles supra-segmentales et syntaxiques. Il s'agit de:

- la résolution des ambiguïtés d'homographies hétérophones et homophones, e.a.
  - le, les, un: déterminant ou pronom
  - porte, reporter, cause, président: nom ou verbe
  - entre, contre: préposition ou verbe
- la détermination des groupes syntaxiques (important pour la synthèse prosodique)
- ajustements phonétiques (élision du *e* muet, assimilation consonantique, dégémination, liaison, pronociation des mots hétérophones).

Tous les homographes ont été repérés en manipulant des listes tirées d'un lexique électronique [7]. Ainsi nous avons répertorié 2750 formes ambigües, dont 66 hétérophones. Le système français utilise des règles micro-syntaxiques pour analyser le contexte local de ces homographes. Par exemple, pour décider de la nature grammaticale et de la prononciation des hétérophones en *-tions*, une des règles essaie de détecter le pronom *nous* dans le contexte immédiat.

Une seconde analyse syntaxique regroupe les mots en syntagmes simples (Groupe Nominal, Verbal ou Prépositionnel), en utilisant des règles sensibles au contexte. Les résultats de cette analyse sont utilisés après dans la synthèse prosodique.

ENVIRONNEMENT DE DEVELOPPEMENT DEPES

Non seulement la stratégie, mais aussi les outils informatiques jouent un rôle prépondérant dans le développement de notre synthèse de parole multilingue. Nous avons mentionné à plusieurs reprises l'emploi de règles et de lexiques. Pour faciliter le développement, nous avons implémenté un environnement de développement flexible, nommé DEPES [9,11].

Toutes les analyses linguistiques susmentionnées sont réalisées à l'intérieur de ce système. L'environnement DEPES génère automatiquement, à partir de règles linguistiques, un code source Pascal et des structures de données. Le formalisme utilisé est calqué sur celui de la phonologie générative [3], et donc familier aux linguistes qui doivent l'utiliser:

```
|O|  --->  |D|   /X_Y;
```

(O = origine; D = destination; X = contexte gauche; Y = contexte droite; les contextes sont facultatifs.)

Voici, par exemple, quelques règles de conversion pour le français :

```
|s|  --> |z| /[vo]_[vo];
|e|  --> |E| /#[co,(1,)]_r(s)#;
|er| --> |e| /_(s)#;
```

Les termes entre crochets désignent des classes, définissables par le linguiste. La première règle sonorise le *s* intervocalique; notez que des frontières morphologiques, introduites préalablement, entraveront la sonorisation du premier *s* dans *présupposer*, par exemple. La deuxième règle se charge de la prononciation du *e* suivi de *r* ou *rs* finals dans des monosyllabes, tandis que la dernière représente la transcription par défaut de *er*. L'emploi de classes complexes permet en outre de formuler des règles encore plus compactes, telle l'assimilation consonantique :

```
|[occl.sonore]| ---> |[occl.sourde]|
   /_[occl.sourde];
```

Ces classes sont déclarées comme suit

```
occl.sonore = [bdgvzZ];
      .sourde = [ptkfsS];
```

Les règles opèrent sur différentes entités linguistiques, réparties sur différentes couches alignées de la structure de données centrale (en général: phonèmes, graphèmes, catégorie grammaticale, intonation). Voici un exemple de cette structure, à une phase intermédiaire du traitement:

```
|#il E   r$-pAr-ti # vEr  sEt ^r#|
| Il est re par ti   vers 7   h. |
| S  X   R          p    u   N  |
|                H         h   a  |
```

DEVELOPPEMENT DES REGLES POUR MOTS ISOLES

Pour chaque langue, nous disposons d'une base de données contenant des mots avec leur transcription phonétique. En utilisant les règles de conversion, le logiciel d'entraînement compare ces transcriptions phonétiques à celles qu'il génère. Les résultats fournissent un score global et sont comparés aux résultats précédents. Tous les mots incorrects sont répartis dans différents fichiers, dont on peut distiller de nouvelles règles. Ce processus itératif est

appliqué à un nombre croissant de mots. Les erreurs restantes constituent finalement le dictionnaire d'exceptions.

Pour accélérer et faciliter ce développement manuel, nous avons implémenté un système d'apprentissage automatique inductif [11]. Le programme a besoin d'une liste de mots orthographiques, avec leurs transcriptions phonétiques. D'abord, il établit la correspondance entre les graphèmes et les phonèmes des mots, à l'aide de la technique des modèles markoviens cachés [10]. Puis, le procès d'apprentissage proprement dit est entamé. Ce procès inductif itératif établit une liste de règles de prononciation pour chaque lettre de l'alphabet. Les règles apparaissent dans l'ordre dans lequel elles doivent être appliquées: les règles spécifiques en haut de la liste, les règles générales en bas. La dernière règle est toujours indépendante du contexte.

Les résultats de l'apprentissage inductif forment un point de départ idéal pour le développement manuel de règles.

CONCLUSION

Nous avons décrit quelques aspects de la stratégie de synthèse de parole à partir d'un texte quelconque, telle qu'elle est utilisée chez LHS. La méthode de développement et les outils employés ont été présentés et illustrés à l'aide d'exemples repris de notre système français.

BIBLIOGRAPHIE

[1] J. Allen, S. Hunnicutt, et D. Klatt (1987), *From Text to Speech: The MITalk System*, Cambridge: Cambridge University Press.

[2] R. Carlson, et B. Granstrom (1976), "A Text-to-Speech System Based Entirely on Rules," *ICASSP-76, Philadelphia*, pp. 686-688.

[3] N. Chomsky, et M. Halle (1968), *The Sound Pattern of English*, New York, Harper and Row.

[4] S. Hunnicutt (1980), "Grapheme-to-Phoneme Rules: a Review," *Speech Transmission Laboratory, Royal Institute of Technology, Stockholm, QPSR 2-3*, pp. 38-60.

[5] S. Oakey, et R. Cawthorn (1981), "Inductive Learning of Pronunciation Rules by Hypothesis Testing and Correction," *Proceedings of the Seventh International Joint Conference on Artificial Intelligence*, vol. 1, pp. 109- 114.

[6] K. Olive, et M. Liberman (1985), "Text-to-Speech: an overview", JASA, Suppl. 1, p. 78, S6.

[7] G. Pérennou, et M. de Calmès (1987), "BDLEX Lexical Data and Knowledge Base of Spoken and Written French," Proceedings European Conference on Speech Technology, Vol. 1, pp. 393-396.

[8] T. Sejnowski, et C. Rosenberg (1987), "Parallel Networks that Learn to Pronounce English Text," *Complex Systems*, vol. 1, pp. 145-148.

[9] B. Van Coile (1989), "The DEPES Development System for Text-to-Speech Synthesis," *Proceedings ICASSP-89, Glasgow*, Vol. 1, pp. 250-253.

[10] B. Van Coile (1990), "Inductive Learning of Grapheme-to-Phoneme Rules," *Proceedings ICSLP-90, Kobe*, Vol. 2, pp. 765-768.

[11] B. Van Coile (1991), "Inductive Learning of Pronunciation Rules with the DEPES system," *Proceedings ICASSP-91, Toronto*, vol. 2, pp. 745-748.

19es J.E.P. - Bruxelles, 19 au 22 mai 1992.

MODELES AUTOREGRESSIFS VECTORIELS ET RECONNAISSANCE DU LOCUTEUR

C. MONTACIE, J.-L. LE FLOCH & X. RODET

LAFORIA - Université Paris 6, CNRS-URA 1095, 4, place Jussieu, 75252 Paris Cedex 5

Résumé

Nous proposons dans cet article une nouvelle méthode pour la reconnaissance du locuteur. Cette méthode utilise une modélisation autorégressive vectorielle à long terme des paramètres spectraux pour caractériser un locuteur ainsi qu'une distance inter-locuteurs discriminante. Plusieurs distances inter-locuteurs sont présentées et leurs avantages/inconvénients discutés. Un mode d'apprentissage discriminant des modèles autorégressifs vectoriels est proposé.
L'évaluation de cette méthode a été conduite sur la base de données TIMIT enregistrée par des locuteurs coopératifs, et sans imposteur. Une série d'expériences d'identification du locuteur indépendemment du texte est détaillée. Aucune phrase spécifique n'est utilisée au cours de l'apprentissage et le corpus de test est différent de celui d'apprentissage. Ces expériences ont donné d'excellents résultats (i.e., un taux d'identification de 99.3% pour 420 locuteurs) en utilisant une seule phrase à chaque test.

1. INTRODUCTION

La reconnaissance du locuteur fait généralement référence à trois applications voisines que sont la vérification d'une identité prétendue, l'identification du locuteur et la détection des changements de locuteurs au cours d'une conversation. Nos travaux portent sur l'identification du locuteur qui est considérée comme la plus difficile de ces trois tâches. Pour que nos résultats ne dépendent pas du contenu phonétique des phrases d'identification, nous n'avons pas utilisé de phrases spécifiques en phase d'apprentissage. La tâche d'identification devient, par conséquent, plus complexe mais bien plus adaptée à une application réelle.
Nous avons développé un système original fondé sur l'utilisation d'un modèle de l'évolution spectrale, les Modèles Autorégressifs Vectoriels (MAV), et sur une distance de distorsion inter-locuteurs que nous avons appelé la Distance d'Itakura Vectorielle (DIV). Les MAV permettent de décrire les trajectoires suivies par les paramètres d'analyse d'un segment de parole (i.e., phrase ou série de phrases).
Pour l'apprentissage et l'identification deux modes sont décrits : un mode discriminant et un mode non-discriminant. L'apprentissage discriminant nous a amené à introduire les modèles Autorégressifs Vectoriels Discriminants (MAVD). L'identification discriminante utilise une distance d'Itakura Vectorielle Discriminante (DIVD). Le mode discriminant donne les meilleurs résultats mais sa complexité est plus grande.
La phase d'apprentissage consiste à apprendre un modèle (MAV ou MAVD) pour chaque locuteur. Le locuteur identifié est celui dont le modèle est le plus proche, au sens de la distance (DIV ou DIVD), du modèle de la phrase-test. L'identification d'un nouveau locuteur nécessite simplement l'apprentissage de son modèle.

2. LES MODELES AR-VECTORIELS

La modélisation Autorégressive Vectorielle est un outil classique pour traiter les signaux à plusieurs composantes. Elle est utilisée ici pour décrire les trajectoires suivies par des vecteurs d'analyse de la parole.
Soit $\{y_n\}$, (n = 1 ,.., N) une suite de N vecteurs spectraux d'ordre p à moyenne nulle. Son évolution est représentée par un modèle autorégressif vectoriel d'ordre q :

$$y_n = \sum_{i=1}^{q} A_i \, y_{n-i} + e_n,$$

où $\{A_i\}$, (i = 1 ,.., q) sont des matrices p x p et $\{e_n\}$, (n = 1 ,.., N) est une fonction d'excitation représentée par un bruit blanc vectoriel centré de matrice de covariance D. Les coefficients des matrices A_i sont estimés par l'algorithme de Levinson-Whittle-Robinson [1]. Le critère d'erreur à minimiser est la trace de la matrice de

covariance D. Les modèles (MAV) n'ont été utilisés que récemment avec des vecteurs d'analyse de la parole. Appliqués en identification du locuteur [2][3][4][5], ils sont interprétés comme une représentation des capacités articulatoires du locuteur (i.e., vitesse et accélération instantannées des paramètres spectraux). La principale difficulté de cette modélisation (MAV) est l'estimation d'un ordre q optimal.

2.1. Estimation de l'ordre optimal du modèle MAV

L'ordre optimal q_{opt} d'un modèle MAV est difficile à estimer. Nous avons choisi d'utiliser le critère d'Akaike [6] bien que ce critère doive s'appliquer à des signaux monodimensionnels gaussiens. Nous avons choisi d'utiliser un ordre 2. Dans nos expériences en identification du locuteur nous avons remarqué que pour un modèle (MAV) d'ordre supérieur à 2, l'erreur de prédiction ne diminuait pas significativement mais que les taux d'identification diminuaient considérablement. Ces résultats sont cohérents avec d'autres expériences sur la modélisation (MAV) du signal de parole [7]. L'interprétation d'un tel ordre (e.g., 2 or 3) est que le mode de transition des vecteurs d'analyse est assez simple et peut être approximé par un modèle MAV d'ordre faible.

3. DISTANCES INTER-LOCUTEURS

L'identification du locuteur consiste à associer une phrase de test à un locuteur original choisi dans un ensemble de locuteurs appris. Cette identification nécessite la définition d'une distance entre locuteurs. nous présentons trois distance IL1, IL2 et IL3. et développons le calcul d'une quatrième distance (DIVD) plus discriminante.

3.1. Distances inter-locuteurs non discriminantes

Les distance inter-locuteurs choisies sont basées sur la distance d'Itakura [8]. Soient les définitions et les notations suivantes. :
$\{x_n\}$ (n=1,..,M) : M vecteurs spectraux du locuteur X.
$\{A_i\}$ (i=1,..,q) : le modèle (MAV) d'ordre q pour $\{x_n\}$.
$\{y_n\}$ (n=1,..,N) : N vecteurs spectraux du locuteur Y.
$\{B_i\}$ (i =1,..,q) : le modèle (MAV) d'ordre q pour $\{y_n\}$.
$\{e_{XA_n}\}$ (n=1,..,M) : le résiduel des vecteurs $\{x_n\}$ filtrés par le modèle $\{A_i\}$.

$$e_{XA_n} = x_n - \sum_{i=1}^{q} A_i \, x_{n-i}$$

$\{e_{YA_n}\}$ (n=1,..,N) : le résiduel des vecteurs $\{y_n\}$ filtrés par le modèle $\{A_i\}$.

$$e_{YA_n} = y_n - \sum_{i=1}^{q} A_i \, y_{n-i}$$

$\{e_{XB_n}\}$ (n=1,..,M) : le résiduel des vecteurs $\{x_n\}$ filtrés par le modèle $\{B_i\}$.

$$e_{XB_n} = x_n - \sum_{i=1}^{q} B_i \, x_{n-i}$$

$\{e_{YB_n}\}$ (n=1,..,N) : le résiduel des vecteurs $\{y_n\}$ filtrés par le modèle $\{B_i\}$.

$$e_{YB_n} = y_n - \sum_{i=1}^{q} B_i \, y_{n-i}$$

D_{XA}, D_{YA}, D_{XB}, D_{YB} : les matrices de covariance des résiduels (i.e., $\{e_{XA_n}\}$, $\{e_{YA_n}\}$, $\{e_{XB_n}\}$, $\{e_{YB_n}\}$).
Nous avons défini trois distances différentes à partir des matrices de covariance. La première distance IL1 est la distance originale développé par Y. Grenier [2]. La deuxième IL2 a été utilisé pour une identification indépendante du texte sur 168 locuteurs [3]. La troisième distance IL3 est une version symétrique de la distance d'Itakura . Elle est fondée sur le fait qu'un modèle (MAV) est moins bien estimé sur une phrase courte que sur une phrase longue. Cette distance donne de meilleurs résultats.

- $IL1(X,Y) = Tr\,(D_{YA})$
- $IL2(X,Y) = \log(Tr\,(D_{YA} * D_{XA}^{-1}))$
- $IL3(X,Y) = \log(M\,Tr\,(D_{YA} * D_{XA}^{-1}) + N\,Tr\,(D_{XB} * D_{YB}^{-1})) / (M + N)$

La discrimination entre locuteurs n'est cependant prise en compte par aucune de ces distances. Chacune cherche à minimiser la distance d'un locuteur à son propre modèle et non à maximiser la distance des autres locuteurs à ce modèle. Cela nous a amené à définir un écart DIVM entre deux locuteurs au cours de l'apprentissage. Cet écart permet de calculer la discrimination obtenue sur l'ensemble d'apprentissage. Une nouvelle distance inter-locuteurs DIVD améliore cette discrimination.

3.2. Discrimination d'une distance

L'écart DIVM entre deux locuteurs S_{l1} et S_{l2} représente la moyenne des distances DIV des phrases d'apprentissage du locuteur S_{l1} par rapport au modèle appris sur les phrases concaténées du locuteur S_{l2}.

Soient les définitions et les notations suivantes :
Nl : le nombre de locuteurs de base d'apprentissage.
Np : le nombre de phrase d'apprentissage par locuteur.
S_l : le $l^{ème}$ locuteur.
$\{v_{njl}\}$ (n=1,..,N_{jl})(j=1,..,Np)(l=1,..,Nl) : les vecteurs spectraux correspondant à la $j^{ème}$ phrase d'apprentissage du locuteur S_l.
$\{V_{nl}\}$ (n=1,..,M_l)(l=1,..,Nl) : les vecteurs spectraux concaténés du locuteur S_l ($M_l = \sum_{j=1}^{Np} N_{jl}$).
$\{t_n\}$ (n=1,..,P) : les vecteurs spectraux correspondant à la phrase-test d'un locuteur inconnu T.
$\{A_{ijl}\}$ (i=1,..,q)(j=1,..,Np)(l=1,..,Nl) : le modèle (MAV) d'ordre q pour $\{v_{njl}\}$ (i.e., modèle d'une phrase).
$\{B_{il}\}$ (i=1,..,q)(l=1,..,Nl) : le modèle (MAV) d'ordre q

pour $\{V_{nl}\}$ (i.e., modèle d'un locuteur).
$\{e1_{njl}\}$ $(n=1,..,N_{jL}))(j=1,..,Np)(l=1,..,Nl)$: le résiduel des vecteurs $\{v_{njl}\}$ filtrés par le modèle $\{A_{ijl}\}$.
$\{e2_{njl_1l_2}\}$ $(n=1,..,N_{jl_1})(j=1,..,Np)(l_1=1,..,Nl)(l_2=1,..,Nl)$: le résiduel des vecteurs $\{v_{njl_1}\}$ filtrés par le modèle $\{B_{il_2}\}$.
$\{e3_{njl_1l_2}\}$ $(n=1,..,M_l))(j=1,..,Np)(l_1=1,..,Nl)(l_2=1,..,Nl)$: le résiduel des vecteurs $\{V_{nl_1}\}$ filtrés par le modèle $\{A_{ijl_2}\}$.
$\{e4_{nl}\}$ $(n=1,..,M_l)(l=1,..,Nl)$: le résiduel des vecteurs $\{V_{nl}\}$ filtré par le modèle $\{B_{il}\}$.
$D1_{jl}$, $D2_{jl_1l_2}$ $D3_{jl_1l_2}$, $D4_l$: les matrices de covariances des résiduels (i.e., $\{e1_{njl}\}$,$\{e2_{njl_1l}\}$, $\{e3_{njl_1l_2}\}$, $\{e4_{nl}\}$).
$IL3_{jl_1l_2}$: La distance IL3 entre la $j^{ème}$ phrase d'apprentissage du locuteur S_{l_1} et le modèle du locuteur S_{l_2}.

$$IL3_{jl_1l_2} = \log\left((N_{jl_1} Tr(D2_{jl_1l_2} D1_{jl_1}^{-1}) + M_{l_2} Tr(D3_{jl_1l_2} D4_{l_2}^{-1})) / (N_{jl_1} + M_l) \right)$$

De la même façon nous pourrions définir $IL1_{jl_1l_2}$ et $IL2_{jl_1l_2}$ avec la distance IL1 (IL2 respectivement).

L'écart DIVM entre un locuteur S_{l_1} et un locuteur S_{l_2} sur l'ensemble d'apprentissage, est alors défini par :

$$DIVM(S_{l_1}, S_{l_2}) = \sum_{j=1}^{Np} IL3_{l_1l_2} / Np.$$

L'écart DIVM intra-locuteur (i.e., $l_1 = l_2$) ne peut être nul que si le nombre de phrases d'apprentissage Np est égal à 1. Il mesure la cohérence des phrases d'apprentissage entre elles. La connaissance de ces écarts va nous permettre de développer une distance discriminante à partir de la distance IL3.

3.3. Distance inter-locuteurs discriminante

La distance DIVD doit posséder, pour être discriminante, les propriétés suivantes : l'écart DIVM calculé grâce à cette distance doit être le plus petit possible pour le même locuteur et être le plus grand possible pour des locuteurs distincts. La solution choisie consiste à calculer DIVD comme une combinaison linéaire des distances $\{IL3(S_l,T)\}$ (l=1,..,Nl) entre la phrase test et les modèles des locuteurs appris :

$$DIVD(S_l,T) = \sum_{k=1}^{Nl} F_{lk} IL3(S_k,T)$$

où F est la matrice Nl x Nl dont les coefficients F_{lk} sont calculé à partir de la matrice E (Nl x Nl) des écarts ($E_{l_1l_2} = DIVM(S_{l_1}, S_{l_2})$).

$F = J E^{-1}$, avec $J_{l_1l_2} = E_{l_1l_2}$ si $l_1 \neq L_2$, 0 sinon.

Le coût de calcul de la distance DIVD devient élevé quand le nombre de locuteurs Nl augmente. La méthode pour diminuer ce coût est de ne plus calculer F sur l'ensemble des locuteurs mais de recalculer F à chaque test à partir des k plus proches locuteurs de la phrase-test. Cette distance permet une amélioration sensible des taux d'identification. Son coût reste toutefois important. Il est préférable de faire une analyse discriminante a priori (i.e., sur les modèles des locuteurs) au lieu d'effectuer une analyse discriminante a posteriori (i.e., sur les distances inter-locuteurs). Il faut calculer pour cela des modèles Autorégressifs Vectoriels Discriminants (MAVD).

4. MODELES AR-VECTORIELS DISCRIMINANTS

Nous cherchons à calculer des MAV minimisant les distances des locuteurs à leur propre modèle, et maximisant les distances des locuteurs aux modèles des autres locuteurs. Les MAVD sont déterminés par la minimisation du critère C satisfaisant ces deux contraintes.

$$C = \frac{1}{Nl} \sum_{l=1}^{Nl} E_{ll} - \frac{1}{Nl(Nl-1)} \sum_{l_1=1}^{Nl} \sum_{l_2=1}^{Nl} E_{l_1l_2} \text{ avec } l_1 \neq l_2$$

Nous avons tout d'abord utilisé IL1, l'écart le plus simple, pour calculer la matrice des écarts E. Une méthode de gradient [10] est utilisée pour la minimisation du critère C. L'algorithme est initialisée par l'ensemble des coefficients des matrices de prédiction des MAV (non-discriminant). Lors de l'identification nous obtenons des locuteurs absorbants (i.e., locuteurs toujours reconnus au détriment des autres). On s'aperçoit qu'à chaque itération les écarts $\{E_{la}\}$ (l=1,..,Nl et a ≠ l), entre un locuteur absorbant S_a et les autres, ne cessent d'augmenter bien que le critère C diminue.

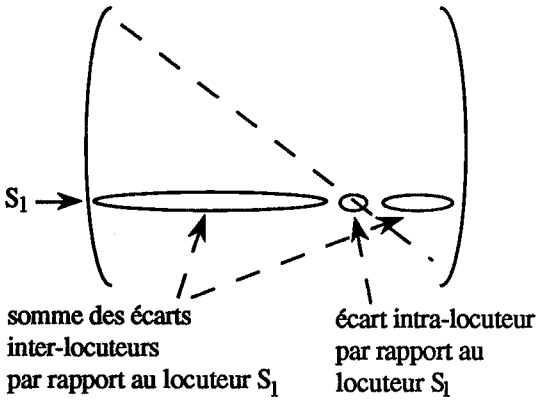


Figure 1. Matrice des écarts E

Pour tout locuteur S_l non-absorbant, ces augmentations sont compatibles avec la minimisation du critère C si la moyenne de la somme des écarts inter-locuteurs par rapport au locuteur S_l augmente plus sensiblement que l'écart intra-locuteur par rapport au locuteur S_l.
La résolution formelle (annulation des dérivées partielles

du critère C par rapport aux coefficients des matrices de prédiction), montre qu'un minimum n'existe pas toujours. Les dérivées partielles sont en effet des fonctions affines des coefficients des matrices de prédiction. Elles possèdent une seule racine, qui correspond quelquefois à un maximum. Cela veut dire que toute autre valeur donnera une valeur du critère plus faible.
Pour résoudre ces problèmes, nous avons alors remplacé la distance IL1 pour calculer la matrice des écarts E par la distance IL3. Dans ce cas, les distances intra-locuteurs (i.e., diagonale de E) augmentent moins vite que les distances inter-locuteur.

5. BASE DE DONNEES

L'identification du locuteur indépendemment du texte est caractérisée par la possibilité de reconnaître un locuteur prononçant n'importe quelle phrase de test. De plus au cours de l'apprentissage, aucune phrase spécifique ne doit être utilisée. Le corpus choisi doit permettre la simulation d'une telle identification.
La base de données utilisée pour nos expériences est la base de données TIMIT. Une description complète de cette base de données peut être trouvée dans [9]. Elle comprend l'enregistrement de 420 locuteurs (i.e., 130 femmes et 290 hommes). Huit "dialectes" régionaux représentent les différents accents de l'anglais américain.
Chaque locuteur prononce 10 phrases. 2 de ces phrases sont prononcées par chacun des locuteurs. Les 8 autres phrases sont différentes pour chaque locuteur. 5 sont des phrases "MIT", les 3 restantes sont des phrases "TI". Les phrases "MIT" (i.e., 450 phrases) sont construites pour avoir une grande variété des contextes phonétiques. Les phrases "TI" (i.e., 1890 phrases) sont extraites d'un corpus de texte écrit. Il n'y a eu qu'une seule session d'enregistrement par locuteur. Aucune étude de de la dérive temporelle d'un locuteur ne pourra donc être effectuée.

6. EXPERIENCES

Les 5 phrases "MIT" sont utilisées pour l'apprentissage d'un modèle MAV* (i.e., MAV ou MAVD). Celui-ci est calculé sur les vecteurs spectraux des 5 phrases concaténées. Le corpus de test est constitué par les 5 phrases restantes. L'identification d'un locuteur est déterminée à partir d'une seule phrase.
Les paramètres utilisés sont des LPCC (Linear Prediction Cepstral Coefficients). Le nombre de coefficients est fixé 20 pour toutes nos expériences sauf pour les modèles discriminants où nous n'en avons utilisé que 8 pour des raisons de complexité de calcul.
Pour identifier un locuteur prononçant une phrase de test, le modèle du test est calculé. Le locuteur reconnu sera celui dont le modèle d'apprentissage est le plus proche au sens de la distance DIV* (i.e., DIV ou DIVD) du modèle du test.
Les expériences d'identification du locuteur ont eu lieu avec des locuteurs coopératifs et sans imposteur.

6.1. Modèles MAV et distances DIV

Pour chacune des distances DIV (i.e., IL1, IL2 et IL3) les résultats sur toute la base de données sont donnés (i.e., taux d'identification sur 420 locuteurs). Des résultats indicatifs sont donnés pour la reconnaissance de N locuteurs. Ces résultats correspondent à une moyenne de 100 taux d'identification obtenus pour 100 tirages au hasard de N locuteurs parmi les 420 locuteurs.

Nombre de locuteurs	20	50	100	300	420
IL1	99.1%	98.1%	97.4%	96.1%	95.3%
IL2	99.7%	99.3%	99.0%	97.9%	97.6%
IL3	99.8%	99.5%	99.3%	98.7%	98.4%

Tableau 1. Identification sur N locuteurs

On peut remarquer (cf. tableau 1.) que le taux d'identification diminue inversement au nombre de locuteurs. Le nombre de locuteurs à identifier est bien un facteur de complexité pour la tâche d'identification du locuteur.

6.2. Modèles MAV et distance DIVD

Le taux d'identification sur toute la base de données est donné en fonction du nombre k de plus proches voisins (kppv) choisi pour le calcul de la matrice F.

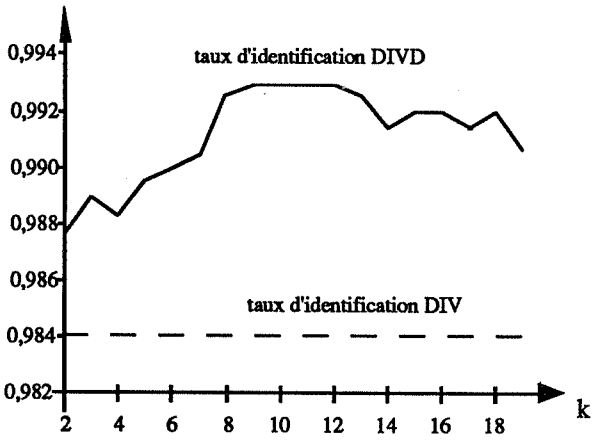


Figure 2. Taux d'identification de DIVD en fonction du nombre k de voisin choisi

On peut remarquer (cf. Figure 2.) que le taux d'identification avec la distance DIVD est toujours supérieur à celui obtenu par la distance DIV (IL3). Ce taux d'identification a un maximum (99.3%) au alentour de 10-ppv. Le tableau 2 donne les 15 erreurs d'identification quand DIVD utilise 9-ppv pour le calcul de F. Pour chaque confusion sont données sur la première ligne les caractéristiques du vrai locuteur, sur la deuxième celles du locuteur reconnu et sur la troisième le texte de la phrase-test.

Nom	Sexe	Dialecte	Age	Taille
JSP	F	1	36	160
PAC	F	7	23	157
That's your headache				
CMR	F	2	34	160
DMS	F	4	32	163
And possessed himself -- how?				
LMC	F	2	22	173
KFB	F	1	??	157
Summertime supper, outside, is a natural				
KMS	F	3	38	173
ELC	F	1	31	163
Perhaps it was right; perhaps it was just.				
DSS	M	2	27	175
GAR	M	7	24	178
He may try to phone us.				
JMG	F	4	32	165
ADG	F	4	26	163
Presently", his water brother said breathlessly.				
REW	F	4	33	163
CKE	F	3	35	173
Bring me the firecrackers.				
SAK	F	4	34	157
ISB	F	7	35	178
Don't ask me to carry an oily rag like that.				
LLL	M	4	26	168
WEM	M	5	29	157
Should she wake him?				
CMB	M	5	37	190
RCW	M	2	27	190
This theory eventually proved inexact.				
MJU	F	6	26	160
SKC	F	3	24	183
Later, you shall know it better.				
PAD	F	6	30	160
KSR	F	7	24	173
Wingman, stay clear, he prayed.				
MAH1	F	7	24	170
DRW	F	6	24	157
Then the telephoning began				
VKB	F	7	24	173
LAG	F	6	24	157
Like his glossy black hair.				
DPB	M	7	26	180
RBC	M	3	25	185
The old shop adage still holds: a good mechanic is usually a bad boss.				

Tableau 2. Confusions sur 2100 tests

On peut remarquer (cf. tableau 2) que les erreurs ne se produisent qu'entre personne du même sexe et pour des phrases généralement courtes (mauvaise estimation des modèles MAV). Par contre, ni le numéro de dialecte, ni l'âge, ni la taille ne semblent être corrélés.

6.3. Modèles MAVD et Distances DIVD

Les expériences avec les modèles discriminants n'ont été menées que sur les 15 couples ayant donné lieu à des confusions dans l'expérience précédente (cf. tableau 2). Une paramétrisation sur 8 LPCC et des modèles MAV* d'ordre 1 sont utilisés. Ces résultats montrent un léger avantage des MAVD (9 reconnus sur 15) par rapport aux MAV (7 reconnus sur 15).

7. CONCLUSIONS

Compte tenu de l'identification indépendante du texte. les résultats montrent que les modèles MAV* permettent la représentation d'une part importante des caractéristiques d'un locuteur. L'identification du locuteur, basé sur des modèles MAV* d'ordre 2, utilise en effet implicitement la vitesse et l'accélération instantannée moyenne des paramétres spectraux qui sont des caractéristiques spécifiques du locuteur (e.g., débit d'élocution).
les modèles MAV* représentent l'effet des caractéristiques du locuteur sur l'évolution des paramètres spectraux. Les résultats obtenus sont très encourageants pour la recherche d'une technique de normalisation de paramètres spectraux au locuteur. Une telle technique pourrait utiliser le résiduel de la modèlisation autorégressive vectorielle à long terme.

REFERENCES

[1] P. WHITTLE : On the Fitting of Multivariate Autoregression and the Approximate Canonical Factorization of a Spectral Density Matrix. Biometrika, Vol. 50, pp. 129-134, 1963.

[2] Y. GRENIER : Utilisation de la prédiction linéaire en reconnaissance et adaptation au locuteur. XI^{ème} JEP, pp. 163-171, Strasbourg, France, 1980.

[3] T. ARTIERES, Y. BENNANI, P. GALLINARI & C. MONTACIE : Connectionist and Conventional Models for Free-Text Talker Identification Tasks. Neuronimes, Nimes, France, 1991.

[4] C.MONTACIE, P. DELEGLISE, F. BIMBOT & M.-J. CARATY : Cinematic Techniques for Speech Processing : Temporal Decomposition and Multivariate Linear Prediction, San Francisco, USA, 1992.

[5] F. BIMBOT, L. MATHAN, A. DE LIMA & G. CHOLLET : Standard and Target driven AR-Vector Models for Speech Analysis and Speaker Recognition. IEEE-ICASSP, San Francisco, USA, 1992.

[6] H. AKAIKE : Information Theory and an Extension of the Maximum Likehood Principle. 2nd Int. Symp. on Informatic Theory. Tsakhadsor, Arménie, URSS, 1971.

[7] O. KAKUSHO & M. YANAGIDA : Hierarchical AR model for Time Varying Speech Signals. IEEE-ICASSP., pp. 1295-1298, Paris, France, 1982.

[8] F. ITAKURA : Minimum Prediction Residual Principle Applied to Speech Recognition. IEEE Trans. ASSP, Vol. 23, pp. 67-72, 1975.

[9] W. FISHER , V. ZUE, J. BERNSTEIN & D. PALLET : An Acoustic-Phonetic Data Base. J. Acoust. Soc. Amer. Suppl. (A), 81, S92, 1986.

[10] M. MINOUX : Programmation mathématique, Dunod.

19es J.E.P. - Bruxelles, 19 au 22 mai 1992.

MODELISATION DYNAMIQUE DES FRICATIVES

CASTELLI Eric ①, SCULLY Celia ②

①INSTITUT DE LA COMMUNICATION PARLEE - GRENOBLE - FRANCE
②DEPARTMENT OF PSYCHOLOGY - UNIVERSITY OF LEEDS - U.K.

Résumé
To produce speech simulations of high quality, we need a good knowledge of temporal coordination between all parts of the respiratory tract. The production of fricative consonants depends primarily on a double articulatory gesture : in the vocal tract a constriction is formed; simultaneously the vocal folds are opened. This double gesture is the cause of (1) an increase in the intra-oral pressure and (2) the production of a frication noise due to the turbulent flow downstream from the constriction. Our work is based on multiple analyses for two subjects. With an analog model of the vocal tract in which an automatic noise source was implemented, analysis-by-synthesis will be used to obtain good aerodynamic and acoustic matches between the natural and the simulated speech, for voiceless fricatives.

INTRODUCTION :

La simulation de parole de bonne qualité nécessite de bien connaître les coordinations temporelles au sein du conduit respiratoire. La production des consonnes fricatives dépend d'un double geste articulatoire : une constriction dans le conduit vocal est réalisée pendant que le sujet ouvre ses cordes vocales. Ce double geste articulatoire produit une augmentation de la pression intra-orale et un bruit de frication dû à l'écoulement turbulent de l'air en sortie de la constriction. Sans nier les effets du couplage avec le conduit vocal sur le comportement de la source vocale [1] [2], nous pensons que la coordination entre constriction et ouverture des cordes vocales est aussi essentielle dans la production des fricatives. En effet, les fricatives ne sont pas seulement caractérisées par leur spectre quasi-statique, mais aussi par les changements spectraux rapides et les changements de sources acoustiques pendant les transitions. Ces changements semblent être importants pour différencier les consonnes /f/ et /θ/ ou les consonnes /v/ et /ð/ par exemple [3]. Nous allons tenter par des mesures en parole naturelle, puis par des simulations temporelles dynamiques de caractériser ces influences.

PAROLE NATURELLE :

Le bruit de frication est fonction de la chute de pression à la constriction, du débit qui la traverse et de sa nature [3]. Les évolutions temporelles de la pression intra-orale et du débit moyen en sortie des lèvres sont dépendantes de la coordination temporelle entre l'ouverture de la glotte et la réalisation de la constriction du conduit vocal. Nous nous sommes alors plus particulièrement intéressés à ces évolutions pendant la production des séquences d'un corpus de parole naturelle comportant les voyelles /a/ et /i/ et les consonnes fricatives /s / et /f/. Pour notre corpus, réalisé dans le cadre d'un projet européen, deux sujets, une femme américaine C.S. et un homme d'origine française P.B., ont enregistré la pression et le débit aérodynamiques dans le conduit vocal (masque de Rothenberg) et le signal de parole (figure n° 1).

Pendant la production des voyelles, le débit est modulé presque entièrement par les cordes vocales, le conduit vocal n'étant pas fortement constrit. La chute de pression entre

la pression subglottique et la pression atmosphérique est localisée à la glotte et la pression intra-orale est pratiquement égale à la pression atmosphérique.

Pendant la production des fricatives, l'évolution rapide de la géométrie du conduit vocal (fermeture par une constriction localisée) et simultanément l'ouverture des cordes vocales compliquent les allures de pression et de débit. Ces deux gestes ont une action conjuguée sur la pression comprise entre la glotte/poumons et la constriction de la consonne, elle augmente et son allure prend celle d'une parabole inversée, pour les fricatives non voisées.

Les deux gestes articulatoires ont par contre des effets opposés sur le débit. L'ouverture des cordes vocales tend à augmenter celui-ci mais ce même débit est fortement diminué par les turbulences au niveau du pincement du conduit. Ces deux effets antagonistes expliquent les pics de débit que l'on observe sur les évolutions mesurées figures n° 1. Il y a diminution de la "résistance" aérodynamique de Ag combinée avec une augmentation de celle de Ac. Les amplitudes respectives des pics U1 et U2 et de la vallée Um qui les sépare dépendent de la nature de la fricative, des voyelles coarticulées avec la consonne (le pic de débit placé après ou avant un /a/ est plus important que celui placé après ou avant un /i/), enfin, elles semblent être très dépendantes du timing entre l'évolution temporelle des cordes vocales et celle de la constriction : le premier pic de débit est dans tous les cas plus important que le deuxième, même dans les séquences où la voyelle est identique (/aCa/ ou /iCi/) [4].

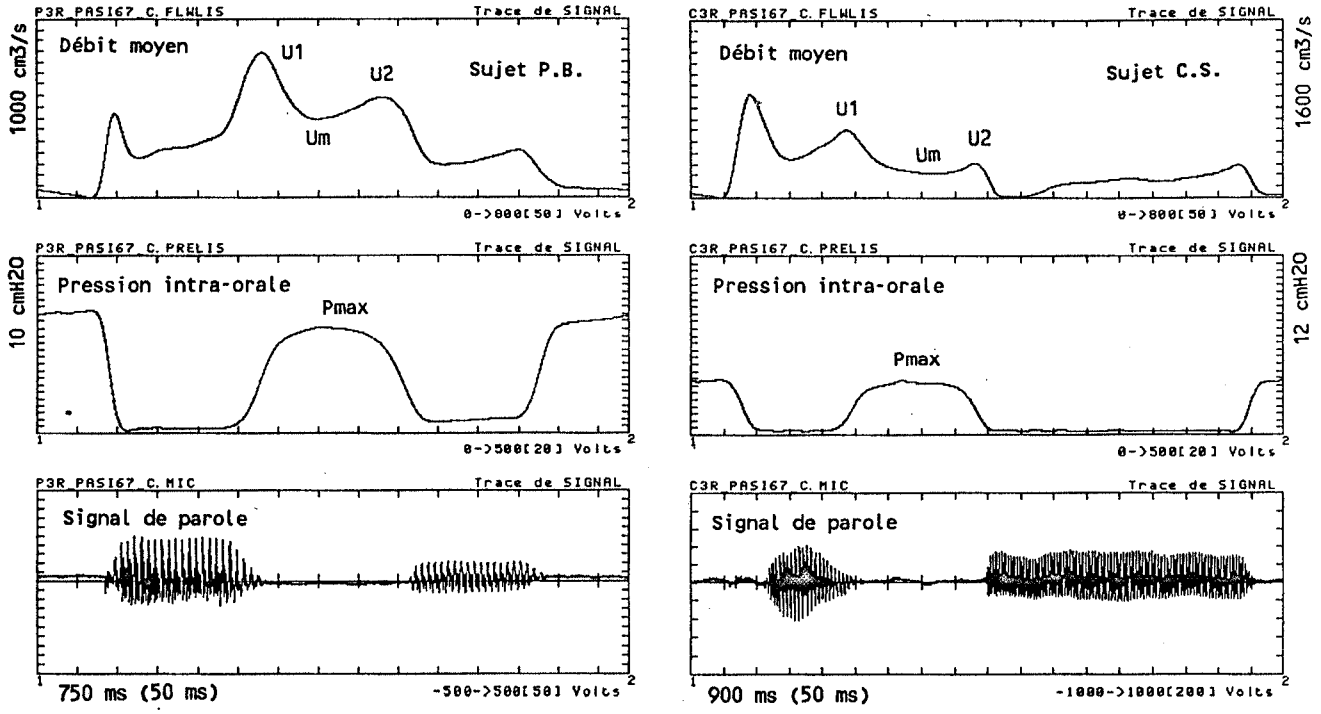


Figure n° 1. Signal, débit moyen et pression intra-orale pour la séquence naturelle /asi/.

MODELISATION :

Le son synthétique est produit par un modèle temporel de simulation de parole, analogue du conduit vocal [5]. Notre modèle, appelé S.I.M.O.N.D., est formé de trois grandes parties : (1) une représentation de l'appareil subglottique (trâchée, bronches et poumons) par un conduit simple dont la fonction d'aire est celle proposée par WEIBEL [6] et dont les résonances mesurées ont des valeurs d'environ 640 Hz, 1400 Hz et 2100 Hz [7] ; (2) le modèle mécanique dit "à deux masses" simulant les cordes vocales [8], amélioré par une commande supplémentaire Ag0 de l'aire d'ouverture au repos des cordes vocales ; (3) une simulation du conduit vocal qui peut être complexe pour accepter éventuellement le conduit nasal et des sinus

[9]. Le couplage entre les cavités subglottiques et supraglottiques et le modèle à deux masses des cordes vocales est amélioré [10]. Les pertes par vibrations des parois, viscosité et chaleur et les effets de rayonnement aux lèvres ou aux narines sont calculés. Les phénomènes aérodynamiques (comme la compliance des parois) sont pris en compte dans S.I.M.O.N.D. et, afin de produire des consonnes fricatives, une source de bruit a été implémentée. La détection du lieu Xc et de l'aire Ac de la constriction se fait automatiquement par détermination de l'impédance moitié de la zone de constriction autour de l'aire minimum du conduit vocal, et la source génère, si besoin, un bruit de frication à pente spectrale variable dont l'amplitude Sp dépend de l'aire Ac et de la chute de pression ΔP à travers la constriction suivant la formule :

$$Sp = k \cdot \Delta P^{p} \cdot Ac^{q}$$

(où k est une constante empirique, p=1.43 et q=0.29 pour /s/, p=0.77et q=0.07 pour /f/ avec le sujet P.B. [11]).

Le bruit ainsi calculé est injecté alors à la position estimée des dents pour la fricative /s/ et aux lèvres pour la fricative /f/.

Afin de piloter ce modèle temporel de manière proche de la parole naturelle, nous l'avons couplé au modèle articulatoire à 7 paramètres de MAEDA [12]. Notre modèle complet, dont les commandes sont alors les paramètres de MAEDA augmentés des commandes du modèle à deux masses des cordes vocales, Ps, Q et Ag0, produit, outre le signal de parole synthétique, le débit glottique, l'aire d'ouverture de la glotte Ag0, l'aire de la constriction Ac, le débit moyen aux lèvres et la pression intra-orale. Les allures des 10 paramètres ont été déterminées afin de reproduire les chemins dynamiques mesurés sur le corpus de parole naturelle [13].

Pour évaluer l'influence de la coordination source/constriction, nous déplaçons l'instant de réalisation de la constriction par rapport à l'instant d'ouverture des cordes vocales. La figure n° 2 nous montre les évolutions dynamiques de l'aire d'ouverture des cordes vocales Ag0 et de l'aire de la constriction Ac, ainsi que des exemples de timing.

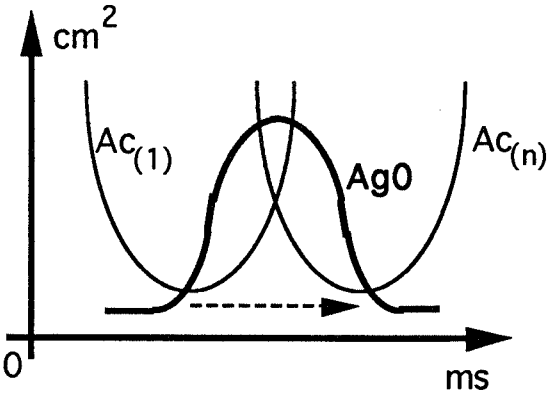


Figure n° 2. Stratégie de variation du timing entre glotte et constriction.

RESULTATS ET DISCUSSION :

Sur la figure n° 3 sont reportées les évolutions temporelles des 10 commandes du modèle complet pour une séquence /asa/ de 600 ms environ. Les paramètres qui varient le plus au cours de la simulation sont le corps de la langue (basculement de la langue entre la voyelle /a/ et la consonne /s/) et la pointe de celle-ci qui doit venir réaliser la constriction contre le palais, juste avant les dents. Les autres paramètres, mâchoire, dos de la langue, hauteur et protrusion des lèvres, larynx, prennent pendant la réalisation de la fricative des valeurs correspondant à une configuration vocalique sensiblement proche du /i/. Parceque la pression subglottique Ps et le facteur masses/tension Q du modèle à deux masses sont constants pendant la simulation, la fréquence fondamentale F0 reste fixe pour les voyelles. La coordination entre glotte et constriction choisie ici permet la production d'un signal synthétique de bonne qualité. Les résultats de la simulation sont présentés figure n° 4.

La figure n° 5 regroupe les allures de débit et de pression produites au cours d'une séquence /isi/ de 600 ms environ où nous avons réglé les différentes coordinations proposées par la figure n° 2. Bien que la consonne /s/ soit coarticulée avec la même voyelle /i/, l'amplitude des pics U1 et U2 de débit dépend bien de la coordination. La vallée Um, quasiment stable en amplitude, est dépendante des pertes dues aux turbulences : elle pourra prendre une valeur plus faible si la durée de la tenue de la fricative est plus importante.

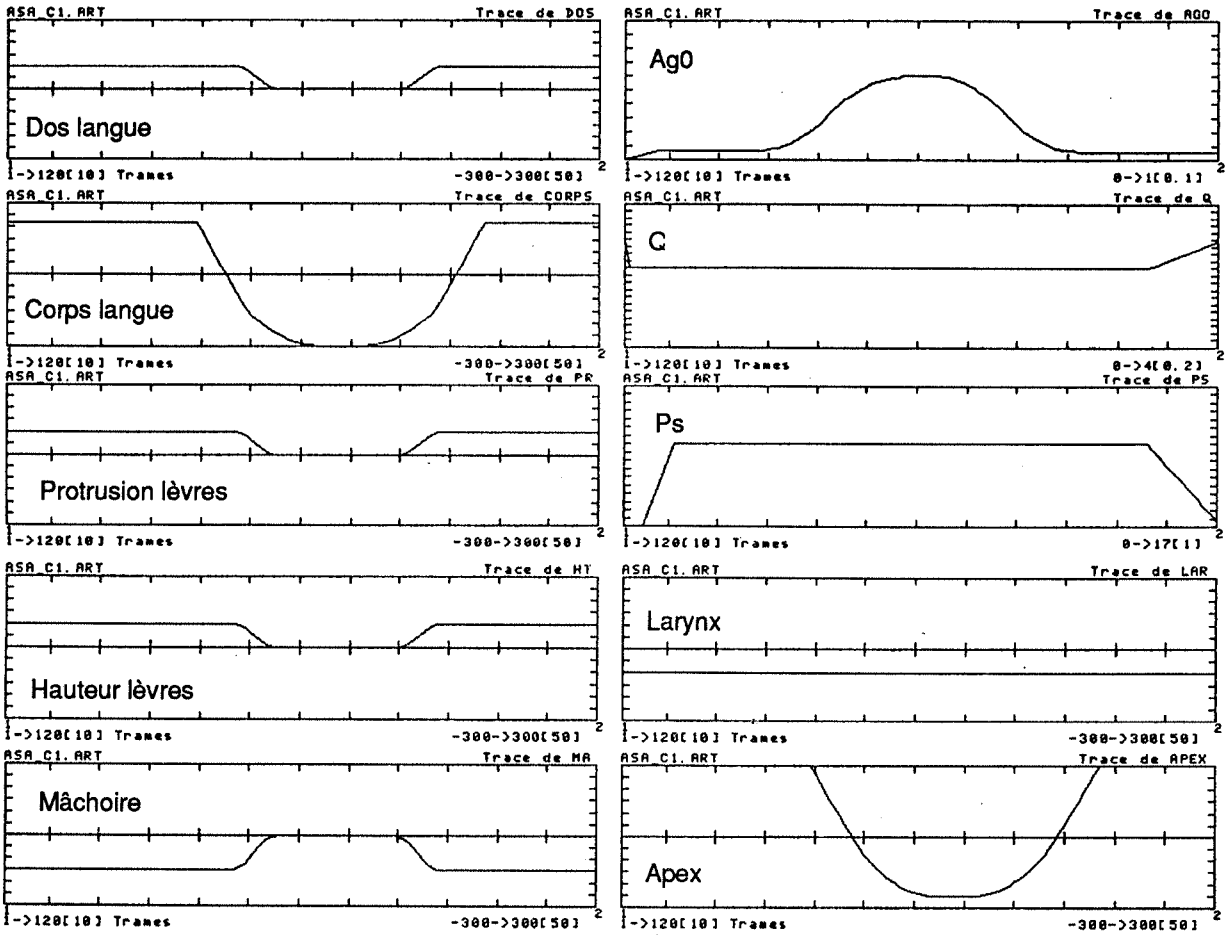


Figure n° 3. Les 10 paramètres de commande pour une simulation de /asa/.

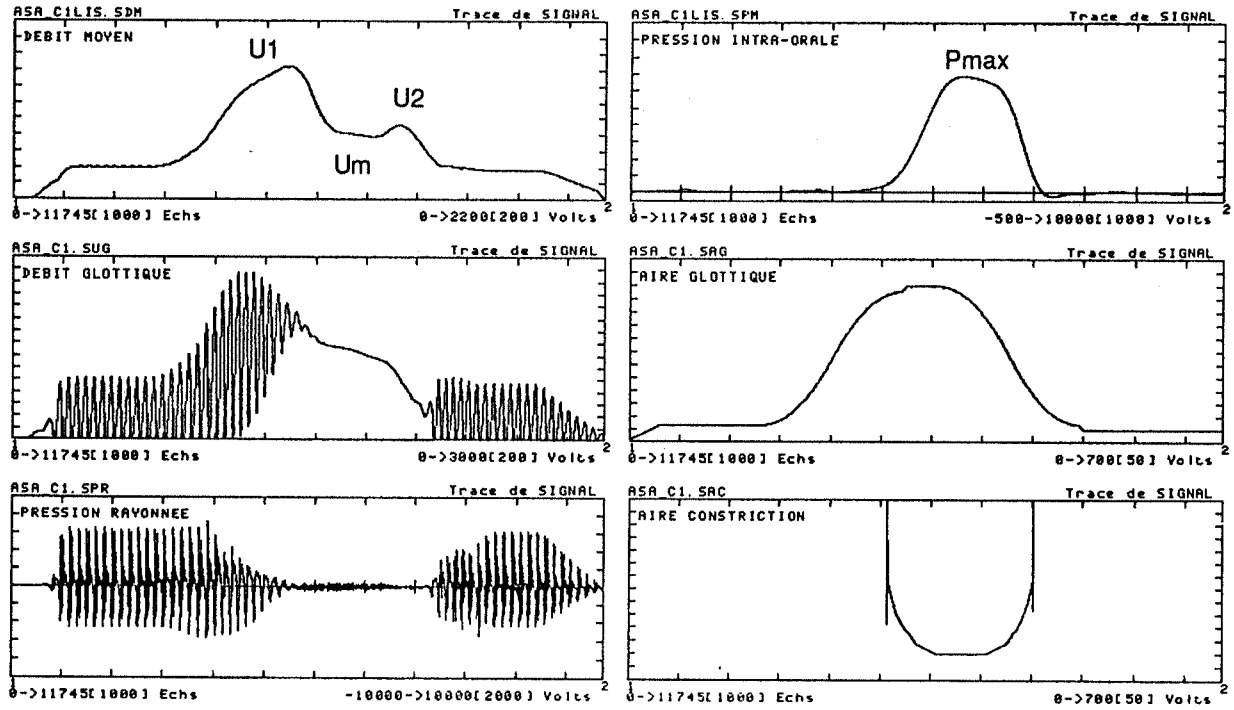


Figure n° 4. Résultats de la simulation de /asa/.

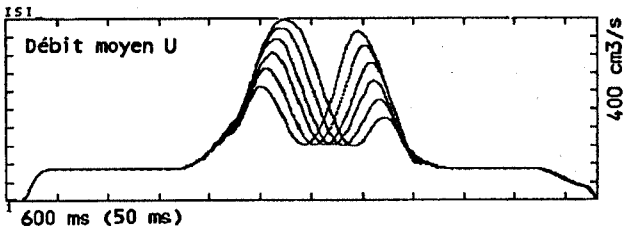


Figure n° 5a. Evolutions du débit pour différentes coordinations

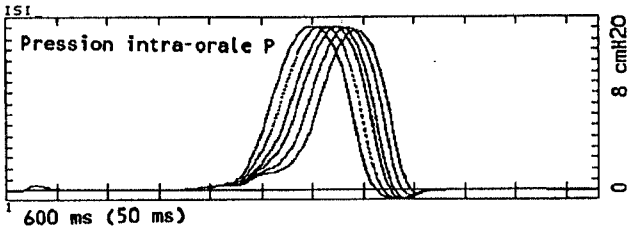


Figure n° 5b. Evolutions de la pression intra-orale pour différentes coordinations

Les allures de pression sont essentiellement stables en amplitude dans nos cas considérés où constriction et ouverture des cordes vocales se produisent presque simultanément. Pour des cas extrêmes, non montrés ici, où la constriction se produirait bien avant ou bien après l'ouverture de la glotte, le maximum Pm serait bien plus petit. Par contre, nous pouvons constater que la position temporelle de ce maximum Pm et la largeur du pic de pression sont bien fonction de la coordination.

Pour produire des pics de débit proches des mesures, les simulations montrent, sur un graphe où sont reportées simultanément Ag0 et Ac (figure n° 6), que les zones optimales de croisement des courbes doivent se situer aux alentours de Ag0 = 3/4.Ag0max pendant l'ouverture de la glotte et Ag0 = 1/4.Ag0max pendant la fermeture. Ceci correspond à des durées d'ouverture de la glotte et de fermeture de la constriction sensiblement égales mais avec un retard de Acmini par rapport à Ag0max d'environ un quart à un cinquième de la durée de la tenue de la fricative.

Acoustiquement, une fermeture de la constriction *trop en avance temporellement* sur l'ouverture de la glotte produit un bruit de frication pendant la fin du voisement de la première voyelle coarticulée : le début de la fricative est alors perçu voisé. En fin de fricative, le bruit de frication cesse (Ac augmente) alors que les cordes vocales n'ont pas encore recommencé à osciller (Ag0 toujours grand) : un silence est produit entre la fricative et la deuxième voyelle coarticulée. De même, une fermeture *trop en retard* aurait des conséquences opposées : des oscillations fortement diminuées, voire nulles, avant que le bruit de frication ne soit audible d'où production d'un silence entre la voyelle et le début de la fricative ; une fin de fricative voisée car le bruit de constriction serait encore produit malgré la fermeture des cordes vocales et le début des oscillations de la deuxième voyelle.

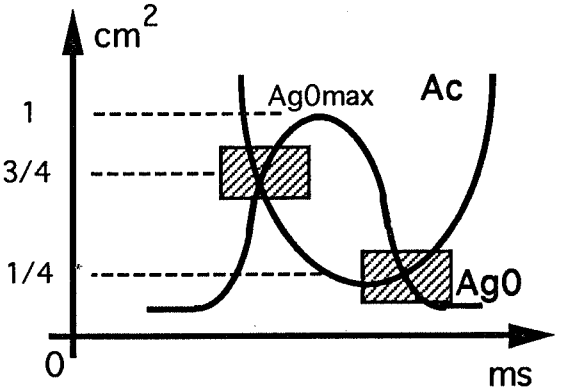


Figure n° 6. Coordination entre glotte et constriction.

CONCLUSIONS :

Les évolutions du débit moyen en sortie des lèvres (et des amplitudes des pics de débit) et de la pression intra-orale pendant la production des fricatives nous ont semblés des paramètres pertinents de notre corpus de parole naturelle pour tenter de caractériser la coordination entre l'ouverture des cordes vocales et la réalisation de la constriction. Nous avons montré par simulation temporelle que les amplitudes des pics de débit et la pression maximale pendant la réalisation des consonnes fricatives non voisées dépendent fortement de ce timing. Nous proposons alors une stratégie dynamique de commande de Ac et Ag0.

Cette première étude doit bien entendue être complétée par des vérifications sur d'autres sujets, par l'étude des autres consonnes fricatives (en particulier /ʃ/ et /θ/) et en vérifiant que notre stratégie s'applique aussi aux fricatives voisées. Une étude des influences spectrales de cette coordination,

sur des spectrogrammes et fonctions de transfert, est envisagée.

Nous avons montré que notre modèle temporel de simulation de la parole produisait de la parole synthétique de bonne qualité, et que son couplage avec le modèle articulatoire de MAEDA permettait de le commander simplement et avec réalisme. Ce modèle complet produisant, outre le signal de parole, des signaux de pression, débit moyen, débit glottique, bruit de frication.... s'avère être un outil d'analyse-synthèse en production très pratique à utiliser. Nous espérons à l'avenir synthétiser les autres types de consonnes et produire de la parole continue.

REFERENCES :

[1] BICKLEY C.A. & STEVENS K.N. (1986) "Effects of a vocal-tract constriction on the glottal source : experimental and modelling studies."Journal of Phonetics n°14, 373-382.

[2] LÖFQVIST A. & YOSHIOKA H. (1991) "Intrasegmental timing : laryngeal-oral coordination in voiceless consonant production." Speech Communication, Vol.3, 279-289.

[3] STEVENS K.N. (1971) "Airflow and Turbulence Noise for fricative and stop consonants : static considerations" J. Acous. Soc. Am.. 50, 1180-1192.

[4] CASTELLI E. & SCULLY C. (1991) "Mécanismes de production des consonnes fricatives : coordination entre glotte et constriction du conduit vocal." 2ème Congres Français d'acoustique, Arcachon, à paraître.

[5] KELLY J.R. & LOCHBAUM C. (1962) " Speech-synthesis", Speech Communication Seminar, Stockholm, 127-130.

[6] WEIBEL E.R. (1963) "Morphometry or the Human Lung" Springer-Verlag, Berlin.

[7] ISHIZAKA K., MATSUDAIRA M. & KANETO T. (1976) "Input Acoustic-Impedance Measurements of the Subglottal System." J. Acoust. Soc. Am., vol. 60, N°1, 190-197.

[8] ISHIZAKA K. & FLANAGAN J.L. (1972) "Synthesis of Voiced Sounds from a Two-Mass Model of the Vocal Cords." B.S.T.J., 51, 1233-1268.

[9] CASTELLI E. (1989) "Caractérisation acoustique des voyelles nasales du français. Mesures, modélisation et simulation temporelle." Thèse Dr. N.R. I.N.P. Grenoble, 181 p.

[10] TRINH VAN L., GUERIN B. & CASTELLI E. (1991) "Couplage entre le modèle à deux masses et un modèle analogue du conduit vocal à réflexion : théorie et implantation." Proceed. of the XIIth International Congress of Phonetic Sciences, Aix en Provence, vol.2, 502-505.

[11] BADIN P. (1989) "Acoutics of voiceless fricatives : production theory and data" STL-QPSR vol.3, 33-55.

[12] MAEDA S. (1990) "Compensatory Articulation during speech : evidence from the analysis and synthesis of vocal tract shapes using an articulatory model" in Speech Production and Speech modeling, W.J. Hardcastle & A. Marchal, eds., 131-149.

[13] SCULLY C. , GEORGES E. & CASTELLI E. (1991) "Fricative consonants and their articulatory trajectories" Proceed. of the XIIth International Congress of Phonetic Sciences, Aix en Provence, vol.3, 58-61.

19es J.E.P. - Bruxelles, 19 au 22 mai 1992.

Mise en oeuvre de phrases arabes phonétiquement équilibrées

par Malika BOUDRAA*, Bachir BOUDRAA*, Bernard GUERIN**

* LCP, Institut d'Electronique, USTHB, BP32 El-ALIA, ALGER
** ICP, 46, Avenue Félix Viallet, 38031, GRENOBLE

Résumé

Cet article résume une méthodologie de mise en oeuvre de phrases arabes phonétiquement équilibrées. La réalisation de ces phrases est basée sur des études statistiques effectuées par A.H.MOUSSA et M.MRAYATI portant sur les racines et les mots de la langue arabe. Une étude similaire a été faite par P.COMBESCURE pour la langue française. Les phrases réalisées nous servent actuellement aux tests objectifs et subjectifs effectués sur plusieurs travaux développés dans notre laboratoire (codage, synthèse, caractérisation des sons de la langue arabe et de détecteurs de mélodie,...).

Introduction

Notre investigation bibliographique sur l'équilibre phonétique de la langue arabe ne nous a pas permis de disposer de listes de phrases arabes phonétiquement équilibrées; ceci, dans le but de répondre aux besoins d'un grand nombre de travaux effectués dans notre laboratoire. C'est le cas des tests subjectifs sur la qualité de la parole de synthèse et des études statistiques entreprises sur la répartition des coefficients de corrélation partielle et des paramètres de l'excitation (optimisation du codage des paramètres LPC et MPLPC...), dans le cas de la langue arabe. C'est aussi le cas de la caractérisation de plusieurs détecteurs de mélodie (études prosodiques de la langue arabe).

Les travaux de A.H.MOUSSA [1,2] et de M.MRAYATI [3], portant sur des études statistiques sur les racines des mots en langue arabe, nous ont permis d'appliquer la méthodologie de P.COMBESCURE [4] (création de phrases et test par $\chi 2$), pour mettre au point des listes de phrases arabes phonétiquement équilibrées.

Equilibre phonétique en langue arabe

Peu d'études ont été faites dans ce domaine. A.H.MOUSSA a effectué des études statistiques sur les racines et sur les mots en langue arabe. Ces études ont été basées sur 3 différents dictionnaires: *Taj Al-arous, Lissan Al-arab, Al-sihah* ainsi que sur le Saint Coran: *Sourate Al Araf* et quelques autres sourates, portant sur 12829 phonèmes consonantiques et autant de phonèmes vocaliques. Ces études ont conduit aux résultats donnés par le table I. M.MRAYATI a complété cette étude par des statistiques faites sur d'autres dictionnaires: *Jamharatu Al-Lughah, Tahdibu Al-lughah, Al Muhkam, Lissanu Al Arab et Al Kamus Al Muhi:t.*

De la table I, on peut tirer les

V / C	Fetha	Kasra	Dhamma	Alif	Ya	Waw	pause	Total
? ء	3.141	1.910	0.514	0.670	0.047	0.047	0.530	6.859
b ب	1.115	1.660	0.374	0.218	0.218	0.218	0.974	4.777
t ت	2.097	0.967	0.834	0.164	0.094	0.086	0.631	4.873
th ث	0.179	0.031	0.133	0.023	0.023	0.023	0.094	0.506
dj ج	0.499	0.171	0.117	0.203	0.023	0.039	0.265	1.317
H ح	0.592	0.156	0.055	0.133	0.125	0.062	0.327	1.450
X خ	0.444	0.086	0.140	0.125	0.031	00	0.203	1.029
d د	0.460	0.304	0.203	0.148	0.171	0.304	0.616	2.206
ḍ ذ	0.327	0.171	0.062	0.398	0.522	0.086	0.437	2.003
r ر	1.653	0.655	0.390	0.242	0.218	0.514	1.274	4.919
z ز	0.164	0.062	0.008	0.016	0.039	00	0.070	0.359
s س	0.787	0.398	0.195	0.304	0.023	0.140	0.865	2.712
ch ش	0.281	0.062	0.070	0.109	0.008	0.023	0.265	0.818
S ص	0.164	0.133	0.070	0.171	0.031	0.016	0.320	0.905
D ض	0.187	0.179	0.023	0.008	0.008	0.008	0.117	0.530
T ط	0.195	0.055	0.008	0.094	0.047	0.016	0.109	0.524
Ḍ ظ	0.078	0.039	0.094	0.055	0.023	0.008	0.078	0.375
E ع	1.723	0.249	0.156	0.156	0.039	0.257	0.631	3.211
g غ	0.179	0.023	0.047	0.094	0.008	0.023	0.164	0.538
f ف	1.419	0.483	0.171	0.109	0.335	0.078	0.288	2.883
q ق	1.099	0.218	0.335	0.772	0.094	0.148	0.288	2.954
k ك	1.208	0.203	1.520	0.413	0.008	0.140	0.312	3.804
l ل	4.116	1.356	0.429	2.027	0.133	0.522	4.373	12.956
m م	1.590	1.980	0.818	1.38 7	0.281	0.405	4.007	10.468
n ن	3.219	0.390	0.320	1.76 9	0.281	0.421	6.462	12.862
h ه	0.335	0.967	1.918	0.795	0.327	0.468	0.249	5.059
w و	3.531	0.016	0.031	0.203	0.023	0.008	1.130	4.942
j ي	1.691	0.171	0.452	0.546	0.023	0.039	1.239	4.161
Total	32.473	13.095	9.487	11.352	3.203	4.099	26.291	100.0

TABLE I: Fréquences d'occurence des voyelles et des consonnes (%).

constatations suivantes:

-La voyelle ou "*haraka fetha*" (brève et longue) apparait le plus fréquemment: 43.825%.

-La pause ou "*haraka soukoun*" apparait pour 26.291%.

-La voyelle ou "*haraka dhammah*"

représente 13.586%.
-La voyelle ou "*haraka kasrah*" apparait 16.298%.
Les phonèmes consonantiques les plus fréquents en langue arabe sont:
/l/ (13%), /m/ (11%), /?/ (7%), /ħ/,/t/,/r/,/w/ (5%). Nos phrases ont été réalisées en tenant compte de ces fréquences.

Distribution expérimentale:

Pour vérifier que les fréquences des phonèmes réalisés dans notre corpus approchent celles données par A.H.MOUSSA, nous avons utilisé le test d'hypothèse $\chi 2$ de PEARSON, connu pour son efficacité en analyse de données. Celui-ci a été utilisé dans les travaux de P.COBESCURE pour la réalisation de phrases phonétiquement équilibrées en langue française.

En d'autres termes, nous utilisons le test de $\chi 2$ pour déterminer dans quelles mesures les distributions des phonèmes dans nos phrases respectent les distributions données par A.H.MOUSSA. Cette hypothèse sera celle dite Ho pour le test de $\chi 2$.

dans notre corpus.

Nous avons retenu également le seuil 5 de $\chi 2$ pour accepter ou rejeter Ho, comme dans le cas des travaux de P.COMBESCURE. Ceci correspond à une valeur très faible pour notre degré de liberté C-1. A titre d'exemple, considérons les phonèmes /?/, /ħ/ et /s/. Leurs fréquences d'occurence (table I) sont respectivement de 6.859, 5.059 et 2.712. Ceci nous amène à considérer, pour un échantillon du corpus de 100 phonèmes consonantiques et autant vocaliques (diphonèmes du type CV), les fréquences Ni d'occurence de 7,5 et 3, respectivement pour /?/, /ħ/ et /s/.

Par ailleurs, un environnement vocalique du type CV précédent, où C=/?/, /ħ/ ou /s/, doit respecter les fréquences propres de la table II . Soit un total de 6.859, 5.059 et 2.712 respectivement.

Cependant, lors de nos réalisations, nous avons considéré seulement les fréquences ayant une valeur élevée. Pour 100 phonèmes réalisés et pour le cas de /?/ par exemple, nous avons considéré les fréquences d'occurence : 3 pour /?a/, 2 pour /?i/ et 1 pour /?a:/, que nous complétons par un /?u/ ou par un /?w/,selon le sens que l'on désire donner à la phrase et afin d'obtenir une

C \ V	fetha	kasra	dhamma	alif	waw	ya	pause	total
?	3.141	1.910	0.514	0.670	0.047	0.047	0.530	6.859
ħ	0.335	0.967	1.918	0.795	0.327	0.468	0.249	5.059
s	0.787	0.398	0.195	0.304	0.023	0.140	0.865	2.712

TABLE II: Fréquences propres d'occurence des consonnes avec leurs *haraka*

Pour nos réalisations, on estimera la distribution:

$$\chi^2 = \sum_{i=1}^{C} \frac{(Fi - \rho_i)^2}{Fi}$$

où C représente les classes phonologiques de la langue arabe. Fi sont les fréquences données dans la table I et fi les fréquences observées

fréquence totale avoisinant celle donnée par A.H.MOUSSA, c'est à dire 7 dans ce cas.

A la base de ce raisonnement fait à titre d'exemple pour /?/, nous avons élaboré la table III dite de pondération. Dans cette table, on regroupe les différentes réalisations individuelles et globales, pour un échantillon de 104 diphonèmes CV, c'est à dire pour chaque consonne et sa haraka.

C \ V	Fetha ـَـ	Kasra ـِـ	Dhamma ـُـ	Alif ا	Ya ي	Waw و	Pause ـْـ	Total
? ء	3	2	(1)	1	--	--	(1)	7
b ب	1	2	(1)	(1)	(1)	(1)	1	5
t ت	2	1	1	--	--	--	1	5
th ث	(1)	--	(1)	--	--	--	--	1
dj ج	(1)	--	--	--	--	--	(1)	1
H ح	(1)	--	--	--	--	--	(1)	1
X خ	(1)	--	--	--	--	--	(1)	1
d د	1	--	--	--	--	--	1	2
ḍ ذ	(1)	--	--	(1)	(1)	--	1	2
r ر	1+(1)	1	(1)	--	--	1	1+(1)	5
z ز	1	--	--	--	--	--	--	1
s س	1	(1)	--	(1)	--	--	1	3
ch ش	(1)	--	--	--	--	--	(1)	1
S ص	(1)	(1)	--	(1)	--	--	(1)	1
D ض	(1)	(1)	--	--	--	--	(1)	1
T ط	(1)	--	--	(1)	--	--	(1)	1
Ḍ ظ	(1)	--	(1)	--	--	--	(1)	1
E ع	1+(1)	(1)	--	--	--	(1)	1	3
g غ	(1)	--	--	--	--	--	(1)	1
f ف	1+(1)	1	--	--	(1)	--	(1)	3
q ق	1+(1)	(1)	(1)	1	--	--	(1)	3
k ك	1+(1)	(1)	1+(1)	(1)	--	--	(1)	4
l ل	4	1+(1)	(1)	2	--	(1)	4+(1)	13
m م	1+(1)	2	1	1+(1)	--	(1)	4	11
n ن	3	(1)	(1)	2	(1)	(1)	6+(1)	13
H ه	(1)	1	2	1	(1)	(1)	--	5
w و	3+(1)	--	--	(1)	--	--	1+(1)	5
j ي	1+(1)	--	(1)	(1)	--	--	1	4
Total	26+(20)	11+(8)	5+(10)	8+(9)	(5)	1+(6)	23+(17)	104

Table III: Pondération des fréquences d'apparition des consonnes avec leur *haraka*

(.) :diphonème en option sans dépassement du total 104 retenu

Calcul du sous X2 et du X2 global:

La table III montre que la base de notre réalisation doit être une liste de 208 phonèmes organisés sous forme de 104 diphonèmes du type CV. Pour l' exemple

C	Fi	Ni	fi	Sous x**2
? ء	6.859	7	6.730	0.0024
b ب	4.777	5	4.807	0.0002
t ت	4.873	5	4.807	0.0009
th ث	0.506	1	0.961	0.4091
dj ج	1.317	1	0.961	0.0962
H ح	1.450	1	0.961	0.1649
x خ	1.029	1	0.961	0.0045
d د	2.206	2	1.923	0.0363
d ذ	2.003	2	1.923	0.0032
r ر	4.919	5	4.807	0.0026
z ز	0.359	1	0.961	1.0095
s س	2.712	3	2.884	0.0109
ch ش	0.818	1	0.961	0.0250
S ص	0.905	1	0.961	0.0037
D ض	0.530	1	0.961	0.3505
T ط	0.524	1	0.961	0.3644
D ظ	0.375	1	0.961	0.9157
E ع	3.211	3	2.884	0.0333
g غ	0.538	1	0.961	0.3325
f ف	2.883	3	2.884	0.0000
q ق	2.954	3	2.884	0.0017
k ك	3.804	4	3.846	0.0005
l ل	12.956	13	12.500	0.0160
m م	10.468	11	10.576	0.0011
n ن	12.862	13	12.500	0.0101
h ه	5.059	5	4.807	0.0125
w و	4.942	5	4.807	0.0037
j ي	4.161	4	3.846	0.0238
tot	100	104	100	3.8352

TABLE IV: fréqueces théorique **et observée pour une consonne et sa"haraka"**

Fi:fréquence théorique d'apparition.
fi:fréquence d'apparition observée.
Ni:nombre observé de diphonèmes.

de liste de phrases donné en annexe, nous avons calculé les fréquences observées pour chaque phonème consonantique. Nous donnons dans la table IV les fréquences théoriques et les fréquences observées dans notre corpus, ainsi que les sous $\chi 2$ correspondant à chacun des diphonèmes observés et calculés par: $S\chi^2 = \frac{(Fi - fi)^2}{Fi}$

A titre d'exemple, pour le phonème /t/, on doit avoir une fréquence avoisinant 4.873. On observe dans le corpus une fréquence de 4.807. Ceci correspond à un sous $\chi 2$ de 0.0009.

Nous avons suivi la même procédure pour le calcul des différentes fréquences observées. Les résultats obtenus sont regroupés dans la table IV.

Enfin, un $\chi 2$ global a été calculé pour l'ensemble des classes. Pour l'exemple de liste précédent, nous avons calculé le $\chi 2$ total par: $\chi^2_{Tot} = \sum^{28} \frac{(Fi - fi)^2}{Fi}$

Sa valeur est de 3.8352.

Dans cette somme, C vaut 28 et représente le nombre de classes de diphonèmes du type CV (consonne et sa *haraka).*

Le degré de liberté correspondant est de 27. Notons que cette classe de diphonèmes représente en fait les 34 phonèmes de la langue arabe.

Les tables de $\chi 2$ montrent que pour un niveau de signification alpha=0.005, la valeur critique de $\chi^2_{0.005}$ vaut 11.8 pour un degré de liberté de 27. La valeur calculée $\chi 2$ expérimentale étant inférieure à la valeur critique, notre hypothèse est donc acceptée avec une probabilité supérieure à 99.5%.

Cas d'une phrase du corpus:

Considérons le cas de la phrase:

صعد الإمام فوق المنبر

SaEada l?ima:mu fawqa lminbari

Cette phrase contient 15 diphonèmes distribués selon la table V.

c \ V	/a/	/i/	/u/	/a:/	/ °/
S ص	1				
E ع	1				
d د	1				
l ل					2
? ء		1			
m م		1	1	1	
f ف	1				
w و					1
q ق	1				
n ن					1
b ب	1				
r ر		1			

Table 5: Distribution des diphonèmes dans la phrase:

"SaEada l?ima:mu fawqa lminbari"

صعد الإمام فوق المنبر

Enregistrement du corpus:

Les différentes phrases du corpus ont été prononcées par 6 locuteurs algériens parlant correctement la langue arabe (3 locutrices et 3 locuteurs). Chacun des locuteurs a répété 3 fois l'enregistrement. Celui-ci a été effectué à l'ICP de Grenoble, grâce à la station EUROPEC [5] et vérifié à l'aide du logiciel PTS [6].
Pour toutes les phrases, un enregistrement simultané des signaux microphonique et laryngographique a été effectué. Ceci nous permet, entre autres, de valider nos detecteurs de mélodie.

Conclusion:

Ce travail nous a permis de mettre au point une base de données qui sert actuellement à valider plusieurs travaux entrepris au sein de notre laboratoire: codage, synthèse et caractérisations prosodiques des sons de la langue arabe. Les enregistrements simultanés microphonique et laryngographique nous permettent de caractériser plusieurs detecteurs de pitch pour tous les sons de la langue arabe.

Références:

[1] A.H.MOUSSA, "Computer application to the Holly Coran, CV relations", Progress In Cybernetics and systems research, Vol.11, pp. 527-31, 1982.

[2] A.H.MOUSSA, "Computer application to arabic Roots and Arabic Words", Applied Arabic Linguistics and Signal and Information Processing (A.A.L.S.I.P.), Hemisphere publishing corp., 1987.

[3] M. MRAYATI, "Statistical studies of Arabic Language Roots", A.A.L.S.I.P., Hemishere Publishing, 1987.

[4] P.COMBESCURE, "20 listes de 10 phrases phonétiquement équilibrées", Revue d'acoustique, no 56, p.34, 1981.

[5] Logiciel EUROPEC, Greco-prc, Sam 2589,ICP, Grenoble.

[6] J.C. CAEROU, J.M. DOLMAZON, Logiciel PTS, Greco-ICP, Grenoble.

Annexe: une liste de 10 phrases arabes phonétiquement équilibrées:

SaEada l?ima:mu fawqa lminbari
صعد الإمام فوق المنبر

namnama ma:?a ljawma
نمنم ماء اليوم

walen jachfiqa Eanha:
ولن يشفق عنها

la: tansi ?akla burtuqa:laki
لا تنس أكل برتقالك

?uskut ja: xaru:d
اسكت يا خرود!

?a:da:hu zaHfa ramlihi
آذاه زحف رمله

kam biEti thamrana:
كم بعت تمرنا

wa sa?ala hal DagaTa
وسأل هل ضغط

?id Danantahu djanbina:
إذ ظننته جنبنا

lwajlu min nubla:
الويل من نبلى

19es J.E.P. - Bruxelles, 19 au 22 mai 1992.

UN SYSTEME D'EVALUATION OBJECTIVE DE LA DYSPHONIE POUR L'AIDE AU DIAGNOSTIC ET LA REEDUCATION FONCTIONNELLE

Bernard TESTON

Institut de Phonétique d'Aix en Provence

RESUME

Nous avons développé un système d'aide à l'évaluation vocale, centré sur un micro-ordinateur auquel sont associés des capteurs acoustiques et aérodynamiques et des circuits de mesure particuliers qui permettent d'enregistrer les paramètres suivants: Signaux acoustique et électro-glottographique, fréquence fondamentale, intensité et débit d'air buccal. Ces paramètres sont édités pour étudier leurs évolutions temporelles. Différents traitements statistiques et acoustiques leurs sont appliqués, pour donner au clinicien le plus d'informations utiles à l'établissement d'un diagnostic et, au contrôle thérapeutique ou d'une rééducation.

I - INTRODUCTION

L'examen clinique des dysphonies qui, comme le précise HIRANO [1], doit permettre de:
- Evaluer le degré de dysphonie.
- Diagnostiquer la maladie causale.
- Déterminer la gravité de celle ci.
- Déterminer le pronostic (de la dysphonie comme de la maladie causale).
_ Analyser les évolutions de la dysphonie en cours ou aprés un traitement.

Le diagnostic de la maladie causale reste pour l'instant , essentiellement du ressort de l'examen vidéo-laryngo-stroboscopique. Il est toujours accompagné d'une évaluation acoustique de la voix pathologique réalisée par le phoniatre, à partir de la description perceptive de différents facteurs tels que, le souffle, la rugosité, l'asthénie, l'enrouement etc... L'évaluation acoustique auditive, des différents niveaux de disfonctionnement du larynx pendant la production vocale, est toujours entachée par la subjectivité des médecins phoniatres. Malgré plusieurs tentatives de normalisation dont l'échelle G.E.R.B.A.S de l'école japonaise [1] et celle de l'école suédoise [2], cette évaluation, trop conditionnée par l'éducation auditive du praticien, ne permet pas de décrire certains aspects de la pathologie.
Des jurys d'écoute tentent de neutraliser ce facteur mais, cette procédure est trop lourde à mettre en oeuvre pour évaluer une dysphonie ou apprécier l'efficacité d'un traitement.

Depuis quelques années, à l'initiative surtout de l'école japonaise, des méthodes d'évaluation objective de l'importance des dysphonies ont été proposées. Une étude du "voice comittee" rapportée par HIRANO [1] fait état d'un sondage mondial sur l'utilisation de différentes méthodes d'évaluation des dysphonies. Une autre étude, rapportée par FRESNEL-ELBAZ [3], fait le point sur l'exploration fonctionnelle du pharyngo-larynx dans 7 pays européens. Ces deux publications montrent que sur une cinqantaine de méthodes connues d'évaluation des dysphonies,à peine plus d'une dizaine font preuve d'un intéret reconnu pour leur efficacité clinique et leur facilité d'utilisation . Elles sont basées sur la mesure de la stabilité du vibrateur laryngien (jitter) [4], sur la puissance et de la stabilité de l'émission vocale (shimmer), sur l'analyse spectrale du signal vocal [5] et, plus

rarement, sur l'étude du débit d'air et du contrôle pneumo- phonatoire [6]. Si chacun de ces paramètres, est susceptible de fournir des indications précieuses, la simultanéité de leur mesure permet, en combinant leurs informations, de réaliser des analyses multiparamètriques plus efficaces. Une précédente étude sur de nombreux sujets [7], nous a montré le bien fondé de cette méthode. Pour la mener à bien, nous avons utilisé différents appareils de mesure sous la forme d'un système non intégré, d'une grande lourdeur de manipulation, d'un principe proche du système PS-77H de SAWASHIMA [8]. A la demande de médecins phoniatres et orl laryngologues, nous avons réalisé ce système d'évaluation des dysphonies en adaptant pour cet usage, les techniques d'analyse du domaine de la phonétique qui nous sont familliéres.

II - DESCRIPTION DU DISPOSITIF

Il se présente sous la forme d'une station de travail , qui regroupe dans un système modulaire, l'essentiel des techniques d'investigation physiologique, pathologique et clinique du domaine de l'ORL Phoniatrie. Les avantages généraux d'un tel système sont multiples:
-prix de revient plus faible que l'ensemble de tous les appareils dont il concentre les possibilités.
-homogénéité d'utilisation, de fonctionnement et de présentation des résultats, permettant de synthétiser le maximum d'information sur un sujet, dans l'unité de temps et de lieu, sans avoir à utiliser de multiples appareils de présentation variée et de fonctionnement disparate.

La station est constituée par un micro ordinateur auquel sont associés des capteurs acoustiques et aérodynamiques ainsi que des circuits de mesure sous la forme de modules indépendants. Chaque investigation est mise en oeuvre au moyen d'un programme particulier. Il est possible, à partir d'une base de départ, de définir une configuration matérielle et logicielle en fonction des investigations souhaitées.

A - Le micro-ordinateur

C'est un HEWLETT-PACKARD HP 386/20 N avec co- processeur 387, 4 Mo de mémoire.
Un disque dur de 120 Mo, une souris et un écran VGA .
Une carte d'acquisition analogique DATA TRANSLATION DT 2812.
Une carte de jeux JOYSTICK.
Une imprimante couleur HEWLETT-PACKARD PAINTJET ou LASERJET monochrome.

B - Les capteurs

Ils permettent de mesurer le débit d'air inspiré et expiré aux lèvres, et les signaux de parole et de vibration du larynx.

1 - Les capteurs acoustiques

Le signal de parole peut être enregistré :
- avec le microphone contenu dans l'embouchure buccale (AKG C409).
- avec un microphone à electret AKG C525.
- avec un microphone de studio BRUEL et KJAER de la série 4000 (dont l'alimentation est fournie par le système).

- à l'aide d'une entrée ligne pour les signaux enregistrés sur magnétophone (analogique ou DAT).

Les signaux de parole captés par les microphones sont calibrés en niveau à la valeur maximale de 120 dB et à la distance de 30 cm (uniquement pour les microphones AKG 525 et BK 4000).

2 - Le capteur de débit d'air buccal

Il est constitué par un pneumotachographe à grille caractérisé par une dynamique supérieure à 50 dB, un faible volume mort et une bonne linéarit(<2%) [9].
La dynamique de mesure du débit est étalée sur 6 gammes :
5 , 2 , 1 , 0.5 , 0.2 , 0.1 dm3 par seconde (ou litre par seconde).
Le capteurs de débit est équipé d'un système d'ajustement du zéro ainsi que d'un filtre passe-bas à phase linéaire pour éliminer les signaux acoustiques qui se superposent aux signal de débit. Le capteur de débit ainsi que le microphone de mesure sont contenus dans une pièce à main solidaire de l'embouchure buccale. Les oscillations laryngiennes sont mises en évidence au moyen d'un éléctroglottographe (EGG) sur le signal duquel on peut visualiser une synchronisation de stroboscope.

C - Les circuits de mesure

Aux circuits de conditionnement des capteurs sont associés les circuits de mesure suivants:

1 - Le détecteur de fréquence fondamentale

Il permet de mesurer la fréquence instantanée des vibrations du larynx période par période à partir, soit du signal de parole capté par un microphone, soit du signal électroglottographique. Il fonctionne en temps réel, sa dynamique de mesure est étalée sur quatre gammes :
- de 120 à 2000 Hz avec une précision de + ou - 2 Hz
- de 60 à 1000 Hz avec une précision de + ou - 1 Hz
- de 30 à 500 Hz avec une précision de + ou - 0,5 Hz
- de 15 à 250 Hz avec une précision de + ou - 0,25 Hz

2 - Le détecteur d'intensité

C'est un sonomètre : il mesure le logarithme de la valeur efficace du signal de parole. Sa constante de temps d'intégration est de 10 ms (On peut la choisir à 50 ms pour les voix masculines très graves). La largeur de bande est comprise entre 20 Hz et 20 kHz, la pondération subjective normalisée «A» est également disponible pour le phonétogramme. La dynamique du détecteur est de 100 dB sur une seule
gamme de 20 à 120 dB.

3 - le système de télécommande

La commande d'enregistrement est laissée à la disposition du manipulateur, elle permet de commander l'enregistrement des paramètres physiologiques sur le micro ordinateur (au moyen d'une carte JOYSTICK) ainsi que sur des magnétophones (REVOX ou DAT).

4 - Le système d'acquisition

Il est constitué par une carte d'acquisition DT 2812 à 16 entrées.
- Les entrées 1,5,9 et 13 sont réservées pour le signal de parole, avec une largeur de bande de 7.5 kHz.
- Les entrées 3,7,11 et 15 sont réservées pour le signal de l'électroglottographe.
- L'entrée 2 est réservée à la fréquence fondamentale, l'entrée 4 à l'intensité et l'entrée 8 au débit d'air buccal.

D - Présentation matérielle des instruments

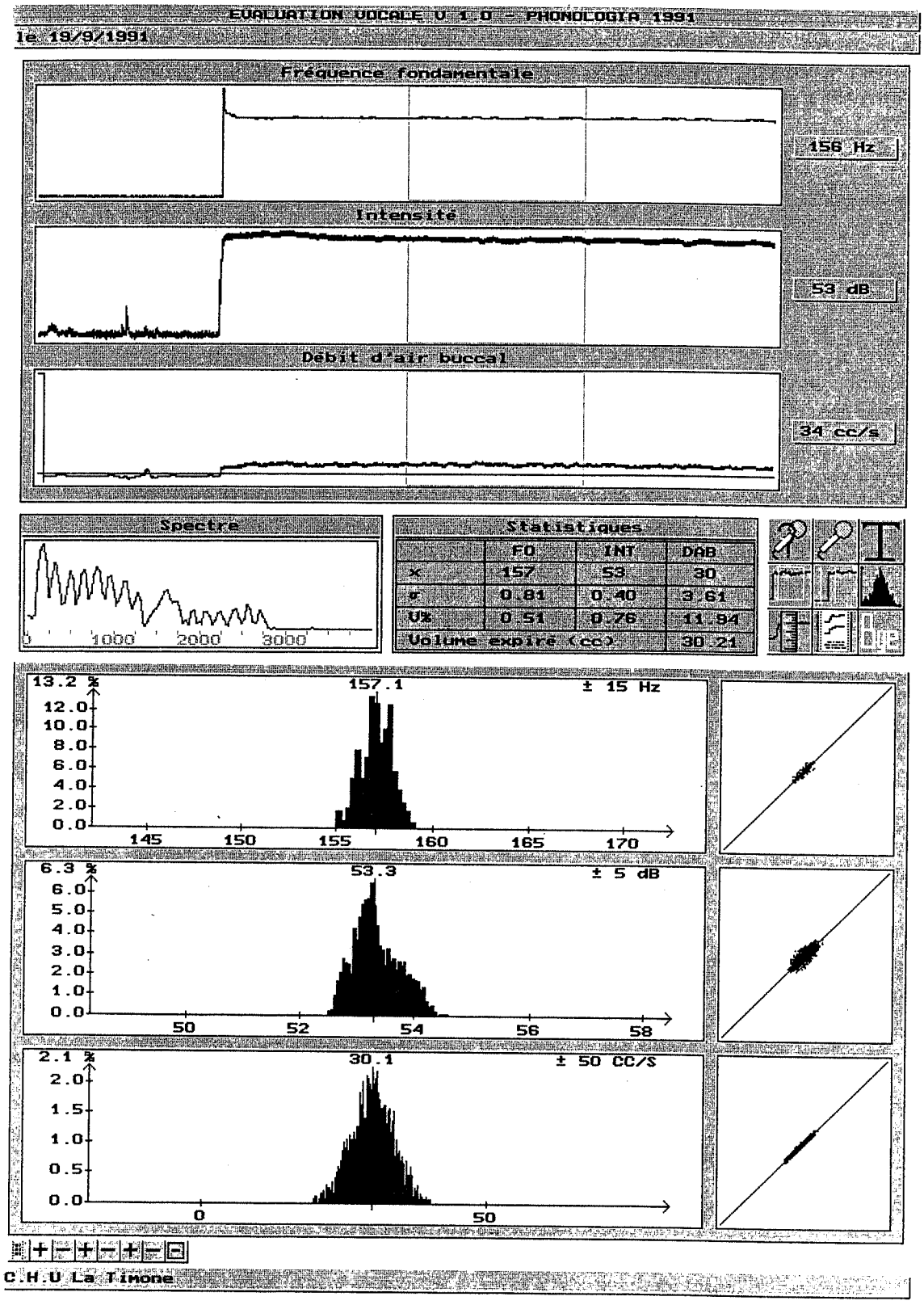
Les capteurs sont contenus dans une pièce à main en acier inoxydable solidaire d'un pied réglable en hauteur. Les conduits de mesure de débit sont réalisés en matière synthétique. Il sont entièrement démontables pour faciliter leur nettoyage. Les embouchures buccales sont réalisés en silicone Rhodorsil. D'un contact agréable, ces pièces peuvent supporter de nombreux cycles de nettoyage (ex Bactinyl). Les circuits de mesure et de conditionnement des capteurs sont contenus dans un coffret au standard «simple europe» (hauteur 15, largeur 46, profondeur 30 cm).

III - CONFIGURATION LOGICIELLE

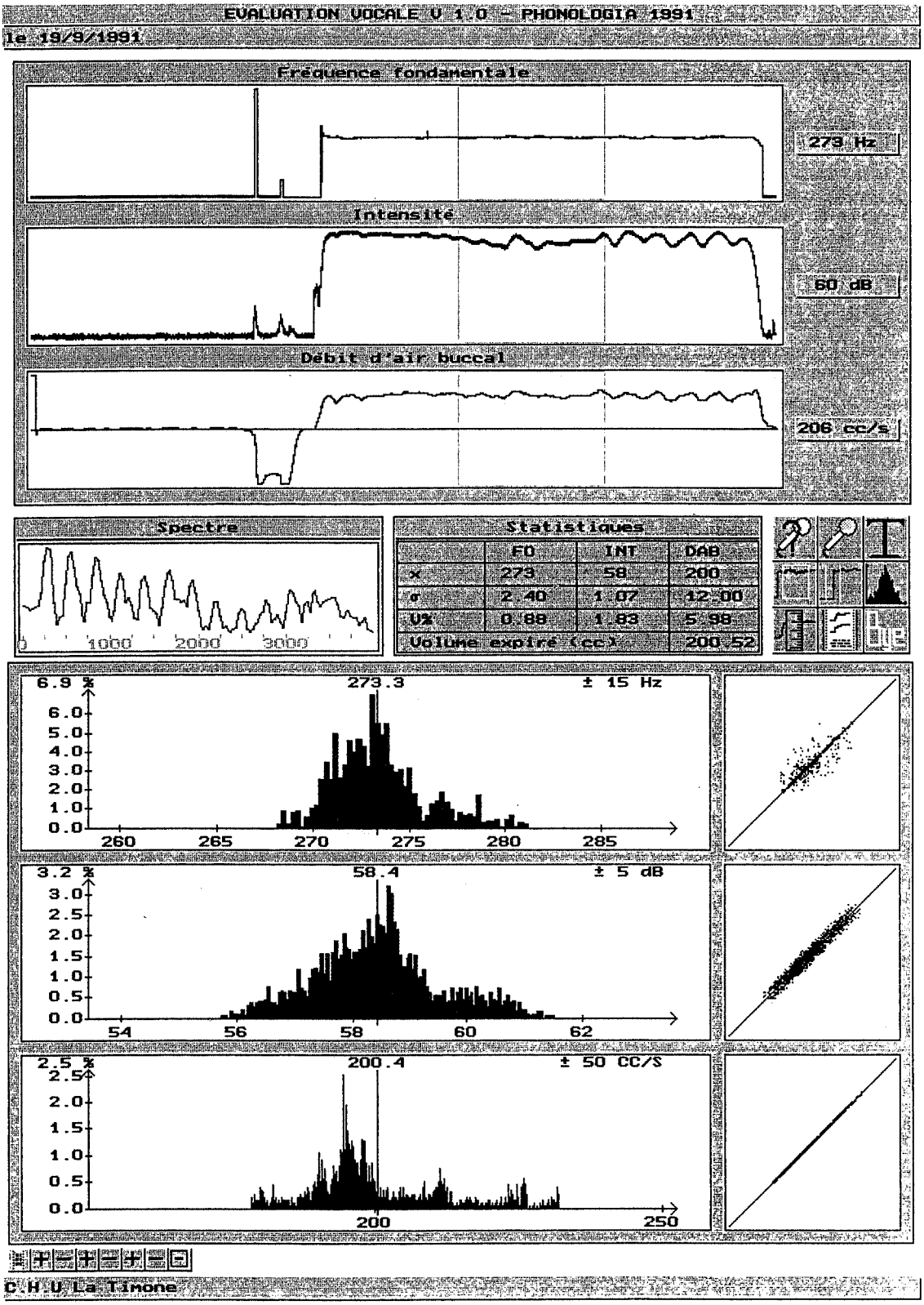
Chaque investigation est mise en oeuvre au moyen d'un programme particulier.

A - Evaluation vocale

Ce programme est destiné à l'évaluation objective de l'intensité d'une dysphonie et à l'aide au suivi
acoustique des pathologies vocales.
-Paramètres enregistrés: Débit d'air buccal, Intensité,
Fréquence fondamentale, Signal micro-phonique.
-Principe: Enregistrement de la voyelle «a» tenue pendant
quelques secondes. On visualise au moyen de quatre fenêtres, les courbes de la variation d'intensité (dB), de mélodie (Hz), et de débit (cc/s) en fonction du temps, ainsi que le spectre fréquentiel du signal acoustique.
Il est possible d'agrandir l'échelle du temps pour étudier des phénomènes fugaces (attaques par exemple).
Une fenêtre d'observation d'une durée d'une seconde est positionnée par le manipulateur sur la partie la plus stable de l'émission vocale. Sur cette durée, les calculs suivants sont effectués:
- Ecart type



Exemple d'analyse vocale sur l'auteur considéré comme un sujet normal. De haut en bas: évolution de la Fo, de l'intensité,du débit d'air buccal, analyse acoustique et traitements statistiques, histogramme de la Fo, de l'intensité et du débit d'air buccal.



Exemple d'analyse vocale sur un sujet féminin dysphonique. On constate une stabilité de Fo satisfaisante ainsi que l'emmergence harmonique par contre, le volume d'air est beaucoup plus important que pour le sujet normal et l'instabilité en fin d'émission vocale de l'intensité et du débit laisse supposer un mauvais contrôle pneumophonique.

- Valeur moyenne
- Coefficient de variation en pourcentage
- Spectre moyen à partir duquel est calculé un indice chiffré d'émergence des harmoniques.

En plus de leurs valeurs numériques, ces résultats statistiques sont représentés sous la forme d'histogram-mes et de nuées bi-dimensionnelles.

IV - CONCLUSION

Les exemples d'utilisation de ce système que nous donnons, ne représentent qu'un aspect de ses possibilités. L'ergonomie de la présentation des résultats bien qu'ayant été mise au point en collaboration avec des cliniciens peut encore être améliorée. La méthode de mesure des paramètres est par contre définitivement figée. Cependant, les traitements , qui ne dépendent que d'un développement logiciel, sont toujours en train d'évoluer . Nous menons actuellement des études comparatives sur les différentes méthodes d'évaluation acoustique de l'émmergence des harmoniques. Nous travaillons également sur différents indices calculés en fonction de plusieurs paramètres proposés par les phoniatres [10] et qui nécessittent une évaluation clinique systématique. Enfin, en ajouttant aux capteurs existants, la mesure du débit d'air nasal, et des pressions intra-orale et sous glottique, nous faisons évoluer la station de travail vers le domaine de la production de la parole dans son entier. Nous lui donnons en particulier, la possibilité d'étudier les rhinolalies congénitales, neuromotices ou acquises ainsi que les dysarthries des différents organes articulateurs.

BIBLIOGRAPHIE

[1] HIRANO, M.(1989).
"Objective evaluation of the human voice Clinical aspects.", **Folia Phoniatrica**, Vol 41, 89-144.

[2] HAMMARBERG, B.,FRITZELL, B., GAUFFIN, J.,SUNDBERG,J. and WEDIN, L., (1980).
"Perceptual and acoustic correlates of abnormal voice qualities.", **Acta Otolaryngologica**, Vol 90, 441-451.

[3] FRESNEL-ELBAZ, E (1988)
"Etat actuel des recherches européennes sur l'exploration fonctionnelle du pharyngo-larynx",**1er Congrés Européen d'ORL**,Paris 26-29 sep. 1988.

[4] SCHOENTGEN, J. (1989)
"Jitter in sustained vowels and isolated sentences produced by dysphonic speakers", **Speech communication**, Vol 8, N°1,61-79.

[5] DEJONCKERE, P. (1986)
"Analyse acoustique de la production vocale. Essai de synthèse dans une optique clinique", **Acta Otorhinolaryngologica Belgica**, Vol 40, N°2, 377-385.

[6] GORDON, M., T., MORTON, F., M. and SIMPSON, I., C. (1978)
"Airflow measurement in diagnostic assesment and treatment of méchanical dysphonia", **Folia Phoniatrica**, Vol 30, N°3, 161-174.

[7] GIOVANNI, A., MOLINES, V., NGUYEN, n. et TESTON, b. (1991)
"L'évaluation objective de la dysphonie", **actes du 12ème congrés international des sciences phonétiques**, Vol5, 274-277.

[8] SAWASHIMA, M. and AOKI, S. (1982).
"A new device for simultaneous recording of airflow rate,vocal pitch and intensity during phonation", **Jpn. J. Med. Instrum.**, Vol 52, 342-345.

[9] TESTON, B. (1988)
"Etude d'un aérophonomètre de grande dynamique et faible constante de temps",**16èmes Journées d'étude sur la parole**,Société Française d'Acoustique, Ham-mammet, 5-9 octobre 1988, 105-108.

[10] ORMEZZANO, Y. (1990).
"Analyse Vocale Immédiate Objective Normalisée", **rapport au congrés de la Société Française de Phonîatrie 1990**, 27 p et 15 ann.

LE GAIN DES LÈVRES : INTELLIGIBILITÉ AUDITIVE ET VISUELLE DE LA PAROLE BRUITÉE EN FRANÇAIS

T. MOHAMADI & C. BENOIT

INSTITUT DE LA COMMUNICATION PARLÉE, U.A. CNRS N° 368,
INPG/ENSERG - Université STENDHAL, BP 25X - 38040 GRENOBLE, FRANCE

Résumé

La perception bimodale permet une meilleure compréhension de la parole que sa seule audition. Dans cet article, nous quantifions l'intelligibilité apportée par le visage du locuteur à la perception de stimuli audio, en fonction de leur dégradation par du bruit blanc. Dix-huit sujets doués d'une audition et d'une vision correctes ont été soumis à un test d'identification à choix fermé portant sur trois voyelles et six consonnes du français, en perception auditive seule d'une part, et audio-visuelle d'autre part. Les taux moyens d'identification correcte fournissent une première mesure de l'amélioration globale obtenue en perception bimodale. La comparaison des matrices de confusion permet en outre de relativiser les effets respectifs de chacune des voyelles et des consonnes. Enfin, nous mesurons l'importance de l'influence contextuelle des trois voyelles [a, i, y] sur les intelligibilités auditive et audio-visuelle des six consonnes [b, v, z, ʒ, ʀ, l] de notre corpus.

1. INTRODUCTION

Avoir la possibilité d'observer le visage d'un locuteur est un avantage considérable pour la compréhension du discours, et tout particulièrement dans un milieu bruité ou réverbérant. Plusieurs travaux ont visé à quantifier l'apport de la vision dans la reconnaissance de la parole dégradée (Sumby & Pollack, 1954; Erber, 1969; Summerfield, 1979). Ces études ont montré que, même pour des sujets normalement entendants et non entraînés à la lecture labiale, les informations faciales augmentent notablement l'intelligibilité de la parole dégradée.

Dans le bruit, les deux modalités visuelle et auditive se complètent lors de la perception de la parole. En effet, ce qui a été dégradé par le bruit dans le spectre de la parole peut être récupéré par la vision des aspects les plus importants de la configuration des lèvres, des dents, et de la langue, qui déterminent le lieu d'articulation de nombreuses consonnes (McGrath et al., 1984).

Sumby et Pollack (1954), puis Erber (1969), ont mesuré le gain en intelligibilité de la perception audio-visuelle (AV) sur l'audition seule (A), en fonction du bruit (S/B) ajouté au signal acoustique. Pour les premiers, ce gain dépend aussi du nombre de mots à identifier : le taux d'intelligibilité de la parole décroît lorsque S/B décroît et/ou le nombre de stimuli augmente, et ce quelle que soit la ou les modalité(s). L'apport de la vision à l'intelligibilité de la parole augmente avec la dégradation acoustique. Pour Erber, cet apport reste stable quand S/B < –30 dB, soit une simple lecture labiale, mais, à –12 dB, le taux de reconnaissance passe de 20% en A à 80% en AV.

MacLeod et Summerfield (1986) ont quantifié l'apport de la perception visuelle comme la différence, en S/B (i.e., en dB), entre les seuils de perception de la parole en A et en AV. Ils ont ainsi mesuré un bénéfice moyen de 11 dB apporté par la lecture faciale.

À ce jour, toutes les expériences connues ont été menées sur la langue anglaise. Nous avons étudié ici le français, en étendant l'analyse de nos résultats à une étude des effets contextuels voyelle-consonne. Ces résultats serviront en outre de référence à des expériences ultérieures pour quantifier et comparer les intelligibilités A et AV de visages parlants synthétiques en cours de développement dans notre laboratoire.

2. MÉTHODE

2.a Corpus

A partir d'un corpus utilisé pour une étude analytique (Benoît et al., 1992), nous avons sélectionné des stimuli de la forme [$V_iC_jV_iC_jV_iz$] (e.g., [iviviz], ou [ababaz], etc.), portés par la phrase interrogative : C'est pas "$V_iC_jV_iC_jV_iz$" ? Notre choix s'est limité à V_i = [i, a, y] et à C_j = [b, v, z, ʒ, ʀ, l]. Les trois voyelles correspondent aux positions extrêmes du mouvement

labial des voyelles ; les quatre premières consonnes sont supposées avoir un effet majeur aux lèvres ; les deux dernières sont supposées neutres, ou jouer un rôle encore mal connu en ce qui concerne la labialité. Dix-huit stimuli différents constituaient donc au total notre test.

2.b stimuli et préparation du test

L'enregistrement et la prise de vue ont été réalisés dans une chambre sourde. Le locuteur, fortement éclairé, a été filmé de face et de profil. Seule, la vue de face, en noir et blanc a été utilisée pour ce test. Les stimuli sélectionnés ont été copiés automatiquement sur une bande vidéo, à l'aide d'un poste de travail spécialisé (Lallouache, 1991) pilotant deux magnétoscopes, l'un en lecture, le second en enregistrement. L'audio a été enregistré sur un seul canal, le deuxième canal étant réservé à l'enregistrement du bruit.

Le bruit de masquage a été produit par un générateur de bruit blanc (spectre plat entre 20 Hz et 20 kHz). Sur le magnétoscope enregistreur, les deux niveaux d'enregistrement audio (canal 1 pour le bruit, canal 2 pour la parole) ont été réglés à l'aide du vumètre, à 0 dB pour le bruit et à -36 dB pour le signal de la parole (la réalisation du phonème [a] de "c'est pas" dans la phrase porteuse servant de référence). Le dispositif expérimental est présenté sur la Figure 1.

Chaque stimulus, son mode de présentation (0 pour A et 1 pour AV), les numéros d'images de début et de fin de sa phrase porteuse, ainsi que le gain de sortie du bruit, ont été préalablement stockés, sur une même ligne dans un fichier descripteur ASCII servant de scénario au dispositif expérimental (voir Figure 2).

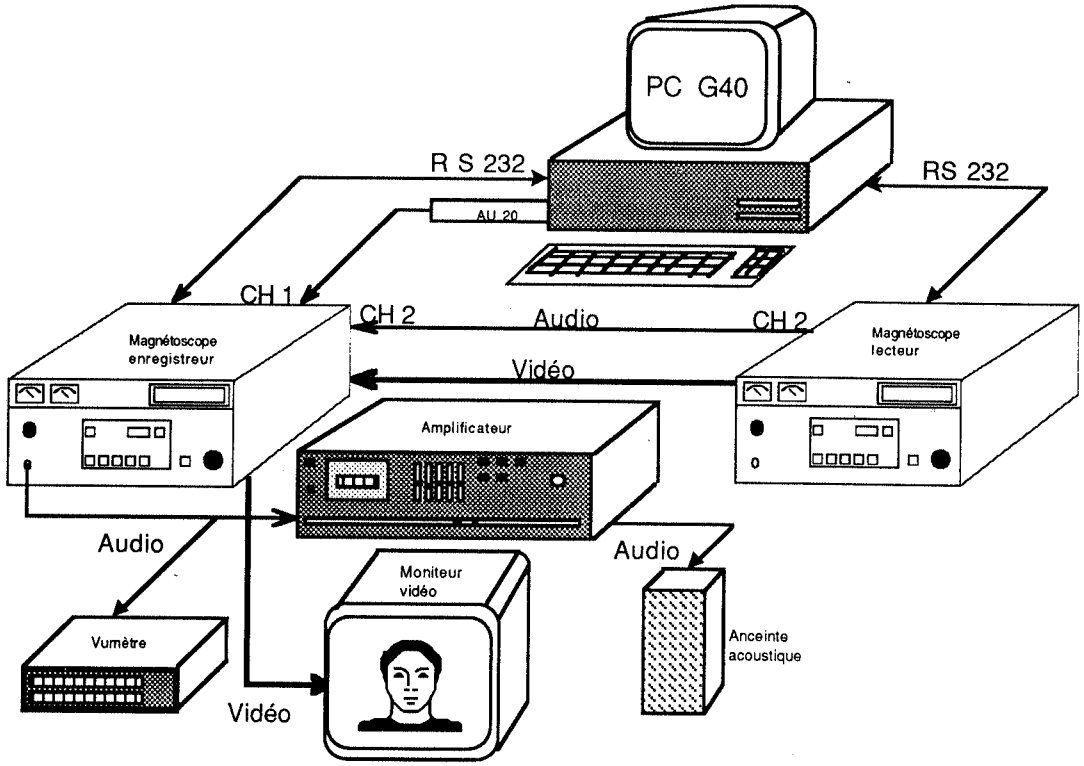


Figure 1. *Dispositif expérimental pour la présentation automatique des stimuli*

Stimulus	Mode A ou AV	trame Début	trame Fin	S/B
IBIBI	1	6468	6534	-36
UVUVU	1	6408	6468	-30
ILILI	1	7984	8040	-24
...	...	...	...	...
ABABA	1	5305	5369	-24

Figure 2. *Exemple de fichier descripteur (partiel) du test en mode AV*

2.c Les sujets

Nous avons retenu 18 sujets français, des deux sexes (11 femmes et 7 hommes), d'âge compris entre 19 et 26 ans (21,5 en moyenne), sans formation particulière en sciences de la parole ni familiarité avec des malentendants. Un audiogramme et un contrôle de l'acuité visuelle de chacun des candidats ont été effectués au début de chaque session. Un seul candidat a été rejeté, pour déficience auditive. Les sujets ont été rémunérés pour leur participation.

2.d Procédure

Le test a été divisé en deux séances individuelles (A et AV) de 24 minutes en chambre sourde. L'ordre de présentation était différent d'un sujet à l'autre (décalage de 5 stimuli à chaque session), avec le même sens de dégradation (du moins au plus dégradé). L'ordre des deux séances a été contrebalancé d'un sujet à l'autre. En face, et à 1,5 m du sujet assis sur une chaise, étaient disposés un moniteur vidéo et une enceinte acoustique.

Les sujets ont reçu des directives, puis ont subi une phase de familiarisation, avec présentation de 5 stimuli, avant chaque séance. Les réponses correspondantes n'ont pas été prises en compte. Cinq feuilles de réponse avec les six consonnes et les trois voyelles disposées sur la même ligne (voir Figure 3) ont été fournies au sujet à chaque séance. Il lui était demandé d'entourer la consonne et la voyelle perçues à chaque stimulus, ou de barrer la ligne correspondante dans les cas d'imperceptibilité (pour éviter l'oubli ou le saut de ligne). Une voix enregistrée invitait le sujet à changer de feuille de réponse tous les 18 stimuli. Un bip sonore précédait de 2 s chaque stimulus. Un délai de 15 s après chaque stimulus (mis à profit pour la recherche automatique de début du stimulus suivant sur la bande) offrait au sujet un temps de réponse confortable.

Nom :_____ Session :___ Audio☐ Vidéo☐ P :___

N°	Consonnes						Voyelles		
1	B	V	Z	J	R	L	I	A	U
2	B	V	Z	J	R	L	I	A	U
17	B	V	Z	J	R	L	I	A	U
18	B	V	Z	J	R	L	I	A	U

Figure 3. *Vue partielle d'une feuille de réponse*

3. SCORES D'INTELLIGIBILITÉ

3.a Résultats globaux

Les réponses de chaque sujet ont été comparées au fichier de référence correspondant. Dans un premier temps, la réponse a été notée fausse en cas d'erreur sur la consonne et/ou sur la voyelle, ou en cas de non-réponse. Le nombre de réponses correctes sur l'ensemble des dix-huit stimuli et des dix-huit sujets, pour chacun des rapports signal à bruit, a été converti en pourcentage. Pour chaque niveau de bruit, l'écart-type a été calculé sur l'ensemble des dix-huit sujets. Les valeurs calculées, ainsi que les deux courbes qui les interpolent sont présentés à la Figure 4.

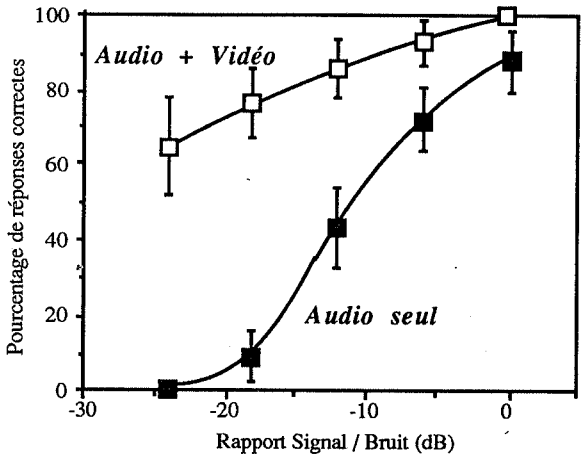


Figure 4. *Taux moyens d'identification de 18 stimuli, en présentation audio seule (courbe du bas), et en audio-visuel (courbe du haut), par 18 sujets, en fonction de la dégradation audio. Les écarts-type des réponses individuelles sont figurés par les segments de droite verticaux sur chaque point de mesure. Les courbes joignant les 5 points ont été obtenues par interpolation.*

Le taux d'identification correcte en audio seul chute brusquement de 72% à 8% pour un écart de rapport signal à bruit de 12 dB (entre S/B = –6 dB et S/B = –18 dB), ce qui est comparable (aux conditions de mesures près) aux observations d'Erber (1969) pour l'anglais. En AV, le taux diminue, pour le même écart, de 93% à 77%. A –24 dB le taux en A est presque nul (0,62%, quand l'aléatoire se situerait à 5,6%), mais il atteint 65% en AV. Ce résultat spectaculaire confirme la facilité avec laquelle les sujets ont pu identifier les stimuli en utilisant quasi-exclusivement la lecture labiale. Nos observations concordent avec celles de Sumby et Pollack (1954) pour le même nombre de mots à identifier : 75% à –30 dB pour 16 mots (à un décalage près d'environ 10 dB en S/B entre nos mesures et les leurs, dû essentiellement au choix de la référence "signal" ; nous y reviendrons plus bas).

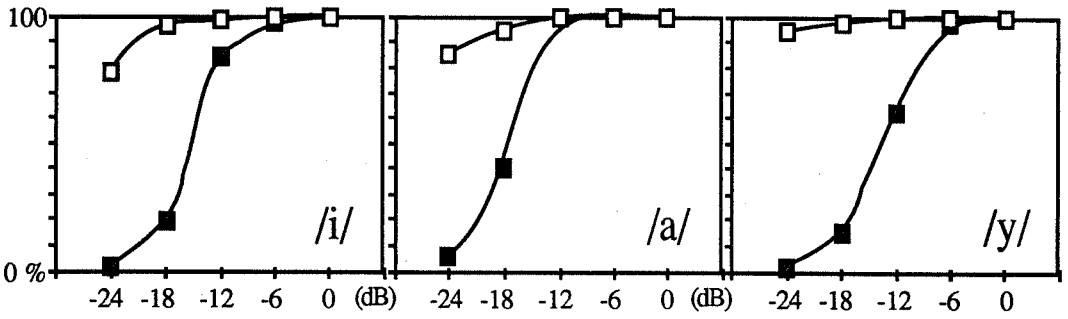


Figure 5. *Intelligibilité A (carrés noirs) et AV (carrés blancs) de [i, a, y], esxprimée en % de voyelles correctement identifiées, tous contextes consonantiques confondus, par 18 sujets.*

3.b Intelligibilité contextuelle

3.2.1 Intelligibilité comparée des voyelles [i, a, y]

Pour permettre une lecture plus détaillée de l'intelligibilité phonétique, la Figure 5 présente les courbes du taux moyen d'intelligibilité de chacune des trois voyelles [i, a, y], tous contextes consonantiques confondus, en fonction du S/B dans les deux conditions de présentation, A et AV.

En perception auditive seule, [a] est plus intelligible que [i], lui-même plus intelligible que [y], toutes choses égales par ailleurs (i.e., tous auditeurs et contextes confondus, cf. à S/B = -18 et -12 dB). Cette observation se justifie par l'intensité intrinsèque des trois voyelles testées : qu'il s'agisse de l'anglais américain, pour lequel Black (1949) a observé une intensité moyenne du [ɑ] de 3,69 dB supérieure à celle du [i], ou du français, pour lequel Guérin et Boë (1978), entre autres, ont observé la même différence entre [a] et [i], l'intensité moyenne de [y] étant comparable à celle du [i]. Or, le S/B calculé ici repose sur l'intensité "signal" de la voyelle [a] de référence mesurée dans "c'est pas...", et non sur l'intensité moyenne du stimulus, ou de la voyelle test. Il faut donc relativiser cette différence d'intelligibilité des trois voyelles par rapport à leur intensité intrinsèque. En effet, les courbes d'intelligibilité acoustique du [a] correspondent grossièrement à celles du [i] ou du [y] à un glissement près de 3 dB environ le long de l'axe des abscisses.

En perception bimodale et sous forte dégradation acoustique (S/B ≤ -18 dB), [y] est mieux identifié que [a], lequel est mieux identifié que [i]. Il faut certainement invoquer ici la spécificité du jeu maxillo-labial nécessaire à la production de ces trois voyelles, et sa robustesse relative aux modifications engendrées par l'environnement consonantique.

3.2.2 Effet du contexte vocalique sur l'intelligibilité des consonnes

L'effet global des trois voyelles sur les intelligibilités A et AV des six consonnes testées est présenté Figure 6, par la projection sur deux axes {intelligibilité A, intelligibilité AV} des différents points obtenus en moyennant sur les cinq conditions de dégradation acoustique les scores de chacune des consonnes, et ce pour chaque contexte vocalique. Chaque consonne est représentée graphiquement par son pourcentage moyen d'identification correcte, tous contextes confondus. De celui-ci partent trois flèches désignant le score obtenu dans chacun des trois contextes vocaliques. Un important effet contextuel est ainsi mis en relief : le contexte vocalique [a] est celui qui favorise nettement l'intelligibilité de toutes les consonnes, tant en A qu'en AV. À l'inverse, [y] est le contexte pessimisant, et ce pour les deux types de présentation (A passe de 30 à 65% et AV de 55 à 90% selon que [z] est présenté en contexte [y] ou [a], par exemple). En revanche, le contexte [i] ne laisse pas apparaître de loi systématique quant à son influence sur l'intelligibilité des consonnes : cette influence est faible, et sa tendance serait à un gain en AV par rapport à la moyenne des autres contextes.

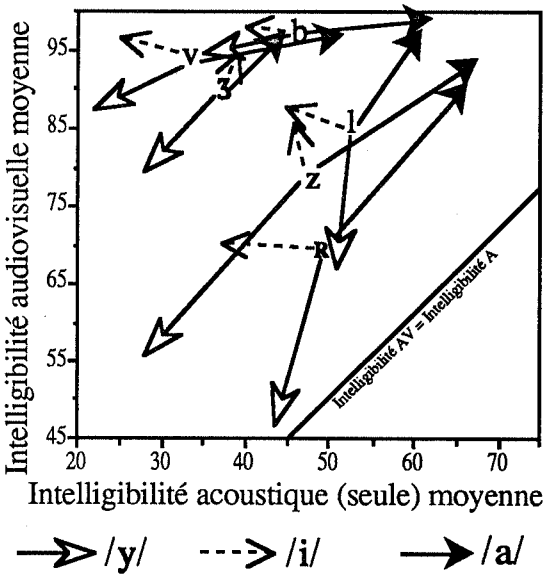


Figure 6. *Effet des voyelles [i], [a], et [y] sur l'intelligibilité des consonnes [b, v, z, ʒ, ʀ, l] exprimée en % de consonnes correctement identifiées (voir texte).*

Si l'on essaye d'extrapoler ces observations à un gain d'intelligibilité apporté par la seule vision du locuteur, et à supposer en première approximation que les deux modalités sont indépendantes l'une de l'autre, il est intéressant de considérer cette projection de la Figure 6 par rapport à la droite {Score AV - Score A = 0} représentée en bas à droite du plan. Tous les points expérimentaux sont situés au-dessus de cette droite, confirmant ainsi le gain d'intelligibilité apporté par la lecture labiale. En revanche, le remarquable effet du contexte [a] vs. [y] a tendance à se situer parallèlement à cette droite, laissant supposer que le gain de la lecture labiale est quasi-indépendant du contexte vocalique pour ce qui concerne l'intelligibilité des consonnes, exceptions faites de [ʀ] et de [l] en contexte [y]. C'est le contexte [i] qui optimise ici l'apport de la lecture labiale des consonnes, puisque les six flèches matérialisant son effet sur la Figure 6 s'éloignent de la droite A = AV.

Il est permis de supposer que l'intensité intrinsèque des voyelles influe sur celle des consonnes environnantes, pour justifier la contribution du contexte [a] à l'apport d'intelligibilité en audio seul. De la même façon, il faut sans doute invoquer la forte coarticulation de [y] (voyelle la plus intelligible en AV) sur les consonnes qui le jouxtent, et la relative plasticité de la forme du [i] (voyelle la moins intelligible en AV) qui dépend souvent davantage du contexte consonantique que l'inverse, pour justifier les différences de gain moyen entre A et AV selon que les consonnes sont perçues en contexte [y] (gain moyen AV – A faible) et [i] (gain moyen AV – A élevé).

4. ANALYSE DES CONFUSIONS

Toutes les réponses des sujets ont été saisies dans une base de données qui nous a servi à établir des matrices de confusion entre les trois voyelles [i, a, y] d'une part, et les six consonnes [b, v, z, ʒ, ʀ, l] d'autre part. Nous avons retenu la condition S/B = –12 dB afin de comparer au mieux les deux modes de perception.

4.a Confusion des voyelles à –12 dB

La Figure 7 présente les deux matrices de confusion des voyelles [i, a, y], à –12 dB, par nos dix-huit sujets, les six contextes consonantiques confondus. Le point d'interrogation figure les non-réponses.

En condition A, la voyelle [a] n'est confondue avec aucune autre. Par contre les voyelles [i] et [y] sont partiellement confondues entre elles, la voyelle [i] ayant tendance à un effet attracteur des réponses (125 percepts pour 108 stimuli) par rapport aux réponses sur la voyelle [y] (79 percepts seulement).

Une désambiguïsation très nette est réalisée en mode AV où les trois voyelles sont perçues correctement quasi-systématiquement.

stimulus \ percept	i	a	y	?
i	91	-	11	6
a	-	107	-	1
y	34	-	68	6
Total :	125	107	79	13

Audio seul

stimulus \ percept	i	a	y	?
i	106	1	-	1
a	-	108	-	-
y	-	-	107	1
Total :	106	109	107	2

Audio + Vidéo

Figure 7. *Matrices de confusion des voyelles [i, a, y], tous contextes consonantiques confondus, par 18 sujets, à S/B = –12 dB. A gauche, confusions auditives seules. A droite, confusions audio-visuelles.*

4.b Confusion des consonnes à -12 dB

La Figure 8 présente les matrices de confusion des six consonnes testées dans les deux conditions A et AV à un niveau de dégradation S/B = –12 dB.

stimulus \ percept A	b	v	z	ʒ	ʀ	l	?
b	27	9	7	3	1	2	5
v	7	20	9	6	3	6	3
z	3	6	31	2	1	3	8
ʒ	5	10	11	12	3	3	10
ʀ	-	5	8	1	32	3	5
l	5	6	1	-	3	33	6
Total :	47	56	67	24	43	50	37

Audio seul

stimulus \ percept AV	b	v	z	ʒ	ʀ	l	?
b	51	2	-	1	-	-	-
v	-	49	1	2	-	-	2
z	-	2	44	5	-	-	3
ʒ	-	-	3	50	1	-	-
ʀ	-	3	1	3	36	10	1
l	-	-	-	-	2	51	1
Total :	51	56	49	61	39	61	7

Audio + Vidéo

Figure 8. *Matrices de confusion des consonnes [b, v, z, ʒ, ʀ, l], tous contextes vocaliques confondus, par 18 sujets, à S/B = –12 dB. A gauche, confusions auditives seules. A droite, confusions audio-visuelles.*

Les consonnes [l, ʀ, z] sont les moins confondues en perception auditive seule. Le score le plus faible est celui de la consonne [ʒ] qui est fréquemment identifiée comme [z] ou [v]. Cet effet n'est pas réciproque : [ʒ] est rarement proposé comme réponse (24 percepts pour 54 stimuli), quand [z] offre, lui, une réponse attractive (67 percepts). Il est intéressant de noter ici la faible confusion [ʀ] / [l] en présentation A.

En présentation AV, la désambiguïsation est encore ici très marquée. Toutes les confusions deux à deux sont en nette diminution, hormis le cas particulier du [ʀ] pour lequel les réponses [ʒ], et surtout [l], sont en augmentation ! En effet, 10 fois sur 54 (au lieu de 3 en mode A), [ʀ] est confondu avec [l]. Comme nous l'avons souligné plus haut, c'est le contexte [y] qui est majoritairement responsable de cette exceptionnelle dégradation de l'intelligibilité au passage de A à AV. La forme maxilo-labiale du [ʀ] et du [l] étant très variable,

la vision du contexte protrus du [y] fait augmenter la confusion [ʀ] -> [l]. Nous n'avons pas d'explication à proposer pour ce phénomène d'attraction non réciproque entre "liquides aux lèvres", sauf à supposer que la réalisation /yʀyʀyz/ testée présente des caractéristiques visuelles particulièrement proches de celles d'un /ylylyz/. Cette interprétation est en accord avec des mesures géométriques antérieures (Benoît et al., 1992), portant sur plusieurs répétitions de /yʀy/ et de /yly/, et qui permettent tout aussi bien d'envisager l'effet attracteur inverse [l] -> [ʀ] avec d'autres réalisations.

4.c Confusions visuelles

La Figure 9 présente les deux matrices de confusion des voyelles (à gauche) et des consonnes (à droite), en présentation AV, à un rapport signal à bruit de -24 dB où aucun son n'est perçu dans le bruit (i.e., en présentation visuelle seule en première approximation).

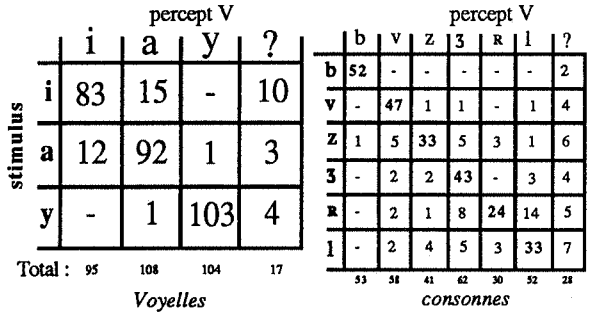


Figure 9. *Matrices de confusion des voyelles [i, a, y], tous contextes consonantiques confondus (à gauche) et des consonnes [b, v, z, 3, ʀ, l], tous contextes vocaliques confondus (à droite), par 18 sujets, à S/B = –24 dB, i.e. pratiquement en mode "video seul".*

Presque toute l'information est apportée ici par la vision du visage du locuteur, en présence d'un bruit de niveau très élevé. On constate que ce sont la voyelle [y] et la consonne [b] qui présentent les formes les plus discriminables. C'est la grande précision du mouvement labial nécessaire à leur production qui est responsable de leur facilité d'identification, quel que soit le contexte : si l'occlusion totale pour le [b] est un truisme, il est bon de rappeler que l'aire intérolabiale du [y] est toujours inférieure à 100 mm^2 (Abry & Boë, 1986). A l'opposé, [i] et [ʀ] sont les moins identifiables, car les plus "élastiques" de notre corpus.

5. CONCLUSION

Les résultats de notre étude pour le français rejoignent globalement ceux obtenus en anglais. Leur analyse détaillée est, par contre, spécifique de la langue : le [ʀ] français a phonétiquement très peu de rapport avec son homologue [ɹ] anglais, par exemple. L'effet du contexte vocalique sur l'intelligibilité des consonnes en parole dégradée, quel que soit le mode de présentation, s'il est confirmé avec d'autres locuteurs français, devra être étudié avec profit par les orthophonistes et les enseignants de français.

Cette étude démontre le remarquable gain d'intelligibilité apporté à la parole acoustiquement dégradée par la présentation du visage d'un locuteur : elle justifie donc l'intérêt technologique d'un développement de synthétiseurs audio-visuels de la parole à partir du texte, tels que celui présenté à ces mêmes Journées d'Etude (Woodward et al., 1992).

Remerciements

Ce travail a été partiellement financé par l'ACCT. Grand merci à Jérôme Zeiliger, Marie-Agnès Cathiard, et Alain Arnal pour leur support technique, ainsi qu'à Christian Abry, Louis-Jean Boë, et Tahar Lallouache pour leurs conseils et leur intérêt pour cette étude...

Références

Abry, C. & **Boë**, L.J. (1986), "Laws for lips", *Speech Communication*, 5, 97-104.

Benoît, C., **Lallouache**, T., **Mohamadi**, T. & **Abry**, C. (1992), "A set of French visemes for visual Speech synthesis", in *Talking Machines : Theories, Models and Applications*, G. Bailly & C. Benoit Eds., Elsevier, North Holland.

Black, J. W. (1949), "Natural frequency, duration, and intensity of vowels in reading", *J. Speech & Hear. Disorders.*, 14, 216-221.

Cathiard, M.A. (1988/1989), "La perception visuelle de la parole: Aperçu de l'état des connaissances". *Bull. Inst. Phonétique Grenoble*, 17-18, 109-193.

Erber, N.P. (1969), "Interaction of audition and vision in the recognition of speech stimuli", *J. Speech & Hear. Res.*, 12, 423-425.

Guérin, B. & **Boë**, L.J. (1978), "Etude d'un indice acoustique des voyelles: la puissance intrinsèque", *9° JEP*, Lannion, 1, 167-176.

Lallouache, T. (1991), "Un poste 'visage-parole" couleur. Acquisition et traitement automatique des contours des lèvres", *Thèse de Doctorat*, INP, Grenoble.

MacLeod, A. & **Summerfield**, Q. (1986), "Quantifying the contribution of vision to speech perception in noise", *British J. of Audiology*, 21, 131-141.

Sumby, W.H. & **Pollack**, I. (1954), "Visual contribution to speech intelligibility in noise", *J. Acoust. Soc. Am.*, 26 (2), 212-215.

Summerfield, Q. (1979), "Use of visual information for phonetic perception", *Phonetica*, 36, 314-331.

McGrath, M., **Summerfield**, Q. & **Brooke**, M. (1984), "Roles of lips and teeth in lipreading vowels", *Proc. of the Inst. of Acous.*, 6, 401-408.

Woodward, P., **Mohamadi**, T., **Benoît**, C., & **Bailly**, G. (1992), "Synthèse à partir du texte d'un visage parlant français", *ce volume*.

19es J.E.P. - Bruxelles, 19 au 22 mai 1992.

EVALUATION PERCEPTIVE D'UN CORPUS DE VOYELLES FRANCAISES EMISES ISOLEMENT PAR PLUSIEURS LOCUTEURS SELON DIVERSES FORCES DE VOIX

Jean-Sylvain LIENARD (*) - Maria-Gabriella DI BENEDETTO (**)

(*) LIMSI-CNRS, Orsay, France
(**) Dipart. INFOCOM, Università La Sapienza, Roma, Italia

Résumé

Le but de cette étude est double. En premier lieu il s'agit de construire une base de données des voyelles du français, validées perceptivement, associant à chaque segment de signal une description linguistique et diagnostique aussi complète que possible. En second lieu il s'agit d'étudier plus précisément, à partir des résultats de l'évaluation perceptive, un descripteur linguistique (l'identité de la voyelle) et deux descripteurs diagnostiques (le genre du locuteur et la force de voix), en eux-mêmes et dans leur relations mutuelles.

I - INTRODUCTION

La présente étude se rapporte à l'élaboration d'une petite base de données sur les voyelles du français prononcées isolément par plusieurs locuteurs avec diverses forces de voix, et à son évaluation perceptive par un groupe d'auditeurs, selon divers critères. Elle est présentée en détail dans [Liénard et Di Benedetto 1992].

L'objectif à long terme est d'aborder les problèmes généraux du traitement automatique de la parole (reconnaissance, synthèse, transmission) selon une perspective qui prenne en compte les aspects non-linguistiques (ou "diagnostiques") du signal [Liénard 1990]. Malheureusement les bases de données existantes se limitent à une description linguistique normative du signal, et sont rarement validées par des tests perceptifs. Pour cette raison nous avons résolu d'enregistrer de nouvelles données et de leur associer une description symbolique validée perceptivement. Dans un premier temps nous avons choisi d'étudier les voyelles, qui sont le support de la voix et de la parole, et dont la reconnaissance est un problème encore non résolu, si l'on considère la variabilité associée à divers locuteurs et à diverses conditions d'élocution.

Cet article décrit d'abord le contenu de la base de données, les conditions d'enregistrement et le protocole d'évaluation. A partir du seul fichier d'évaluation on tente ensuite de définir plus précisément la nature des divers descripteurs mis en jeu, concernant essentiellement l'identité de la voyelle, le genre du locuteur, et la force de voix. On en conclut que les interactions entre ces descripteurs sont nombreuses et importantes, et qu'elles ne peuvent être négligées en traitement automatique de la parole.

II - LE CORPUS ET SON ENREGISTREMENT

La base de données CORENC comporte une série de 12 voyelles françaises, prononcées par 13 locuteurs d'un même groupe familial (6 hommes et 7 femmes, d'âge compris entre 19 et 88 ans) selon trois "forces de voix" (ou styles) différentes, en deux sessions distantes de six mois. L'utilisation de voyelles isolées se justifie par le fait que la plupart des voyelles correspondent en français à un mot lexical (voir par exemple "ah", "a", "euh", "eux", "hi", "y", "oh", "haut", "où", "houx", "hue", "et", "hé", "est", "haie", "en", "an", "on", "un", "hein", etc). Les seules voyelles qui n'apparaissent pas dans cette liste (mais qui apparaîtraient dans des variantes régionales) sont /ɔ/ et /oe/. Dans la préparation de la base de données numérique on a laissé de côté /õe/, /ɔ/ et les deux variantes du /A/.

Le locuteur était assis en un endroit bien défini (dans une pièce d'habitation), la bouche à 30 cm d'un microphone omnidirectionnel. Pour faire varier le style de voix du locuteur on lui a demandé de répéter les voyelles prononcées par l'opérateur, à un niveau sonore adapté à la distance à laquelle se trouvait celui-ci. Ainsi chaque ensemble de trois séries se compose d'une série "N" (voix Normale, interlocuteurs à environ 1,50 m l'un de l'autre), d'une série "P" (situation "Proche", voix plus faible, interlocuteurs à environ 40 cm), d'une série "L" (situation "Lointaine", voix plus forte, interlocuteurs à environ 6 mètres). Les variations de style se traduisent essentiellement par des variations de "force de voix" -- et nous utiliserons indifféremment les deux termes dans cette étude -- mais il ne faudrait pas en déduire

qu'elles se réduisent à de simples variations de niveau sonore. Les nombreux paramètres physiques mis en jeu dans la transmission de cette information (amplitude, mais aussi pente spectrale, position des formants, Fo, texture sonore) feront l'objet d'une étude ultérieure.

Le signal a été numérisé à 10 kHz, le niveau d'entrée sur la carte de numérisation étant maintenu constant (sauf pour les segments "S", correspondant à certains segments de la série "L", voix forte, pour lesquels on a pratiqué une atténuation de 6 dB). Chaque segment a été écouté, visualisé sous forme de spectrogramme, délimité (de -50 ms approximativement avant l'apparition du voisement, à +50 ms après son extinction), et rangé dans un fichier de signal.

III - EVALUATION PERCEPTIVE

A chaque segment de signal du corpus seront associés divers descripteurs symboliques:

- l'identité de la voyelle que l'on a demandé au locuteur de prononcer ("voyelle requise"),
- l'identité de la voyelle perçue majoritairement par un groupe d'auditeurs ("voyelle perçue"),
- l'identité du locuteur qui a effectivement produit le segment considéré ("locuteur réel", ou "locuteur"),
- le genre du locuteur effectif ("genre réel", ou "genre locuteur"),
- le caractère masculin ou féminin de la voix, tel qu'il est perçu ("genre perçu"),
- la "force de voix" qui a été suggérée au locuteur ("style requis"),
- la "force de voix", telle qu'elle a été perçue par le groupe d'auditeurs ("style perçu").

Les segments sont présentés en ordre aléatoire et écoutés par les auditeurs au moyen d'un casque professionnel. Les 40 premiers segments sont répétés à la fin du test de façon à ménager une période d'habituation. L'auditeur écoute chaque segment une seule fois et reporte ses évaluations sur un formulaire, les différentes cases étant notées par deux lettres (p.ex. "aa", "éé", "in", "ee", "eu" etc pour les voyelles, "hh" ou "ff" pour le genre, "pp", "mf" ou "ff" pour le style). Il peut cocher une case spéciale ("??") en cas d'indécision. Il peut régler à sa convenance le niveau d'écoute au début du test. Le test est relativement long (environ 3 heures) et se fait en plusieurs séances. En moyenne le temps séparant l'écoute de deux segments successifs est de l'ordre de 15 secondes. L'évaluation a été faite par 7 sujets. On a ainsi obtenu des taux bruts portant sur 5824 évaluations élémentaires pour chaque descripteur.

En suivant l'évolution de ces taux bruts par tranches de 26 segments du début à la fin du test on n'a mis en évidence qu'un très faible effet d'accoutumance, variable selon les auditeurs, concernant plus les descripteurs diagnostiques que le descripteur phonétique.

On n'a pas inclus dans les résultats définitifs les évaluations de deux auditeurs, l'un parce qu'il était aussi locuteur et l'autre parce que ses résultats se situaient nettement en dehors de la moyenne.

En ce qui concerne le style perçu chaque auditeur n'avait que trois réponses possibles (voix faible, moyenne ou forte). Les réponses des cinq auditeurs sélectionnés ne coincident pas toujours. Pour classer chaque segment dans une des trois catégories une valeur numérique est associée à chaque réponse (soit 0.0 pour une réponse "voix faible", 1.0 pour une réponse "voix moyenne" et 2.0 pour une réponse "voix forte"). A chaque segment est associée la note moyenne comprise entre 0 et 2 et variant par pas de 0.13. La limite entre les catégories (perçues) P et N est définie comme le milieu de l'intervalle séparant les notes moyennes des catégories (requises) P et N. Le même processus est utilisé pour délimiter les catégories (perçues) N et L.

Les segments S ont été utilisés pour vérifier la cohérence des évaluations de style en ce qui concerne une simple différence de niveau sonore. On a constaté que le niveau sonore jouait un rôle dans l'évaluation du style, mais que ce rôle pouvait être considéré comme secondaire pour notre propos: une voix faible, amplifiée électroniquement, ne donne pas une voix forte.

Le résultat de l'évaluation est matérialisé par le fichier descripteur des segments. Il comprend le nom de chaque segment, les trois descripteurs "voyelle requise", "genre locuteur", "style requis", les trois descripteurs perceptifs correspondants, et les notes (3, 4 ou 5) qui indiquent avec quelle majorité ont été évaluées les valeurs des descripteurs "voyelle perçue" et "genre perçu". Cette note n'est pas explicitement donnée pour le descripteur "style perçu", puisqu'elle a servi à déterminer les limites des catégories de style perçu. On a également porté en fin de ligne une évaluation résumée, sous forme de "+" et de "-", qui indiquent pour chaque descripteur si la valeur du descripteur perçu correspond ou non à la valeur du descripteur requis.

Sur l'ensemble des 792 segments de la base le nombre de segments évalués comme corrects sous tous les aspects simultanément (segments "+++") est de 491. Il y a donc 301 segments comportant au moins une erreur, soit 38.0%. Un seul segment a été évalué comme erroné sous tous les aspects ("---"), soit 0.1%. Dans tout ce qui suivra les 72 segments S seront exclus des statistiques.

Le tableau 1 montre que les erreurs phonétiques se répartissent à peu près régulièrement selon le locuteur entre 0 et 30%, ainsi que les erreurs sur le style (entre 17 et 55%) ; mais les erreurs sur le genre du locuteur sont presque uniquement relatives à un seul locuteur (JB). Le locuteur produisant le maximum d'erreurs de style est une personne âgée.

loc	gen	n	%voy	%gen	%sty	pmf
AB	F	36	11.1	0.	38.9	++
AM	F	72	4.1	1.4	31.9	14
CB	F	72	16.7	0.	30.6	-4
JB	F	108	4.6	25.0	16.7	0
MF	F	36	16.7	2.8	55.6	10
SA	F	36	5.6	0.	38.9	-3
SB	F	36	13.9	0.	27.8	-6
BB	H	36	8.3	2.8	22.2	-5
DB	H	36	2.8	0.	25.0	--
JP	H	72	12.5	0.	27.8	0
MB	H	72	6.9	0.	19.4	11
ML	H	36	30.6	0.	27.8	-10
OB	H	72	0.	0.	33.3	-2
ens		720	9.2	4.0	28.6	1

Tableau 1 : erreurs sur l'ensemble du corpus, segments S exclus, par locuteur; n est le nombre de segments fournis par chaque locuteur, pmf est le rapport (plus fort/moins fort), voir partie VI

Le décompte des erreurs phonétiques faites par les groupes de locuteurs masculins et féminins ne fait pas apparaître de différence significative. La comparaison des résultats d'évaluation obtenus dans les mêmes conditions par les 6 locuteurs qui ont pris part aux deux sessions montre une différence légère mais non significative.

IV - DESCRIPTEUR "IDENTITE DE LA VOYELLE"

On trouve dans la littérature scientifique de nombreuses études sur la perception des voyelles. Parmi les principales lignes de force du thème se trouvent l'influence du contexte phonétique, l'influence de la tâche assignée aux auditeurs, et l'influence de l'ordre de présentation des signaux successifs.

Lorsque l'étude est faite sur des voyelles dites isolées, il s'agit très souvent, en fait, de syllabes CVC. Dans une étude classique des voyelles américaines [Peterson and Barney 1952] celles-ci sont prononcées dans un environnement /h-d/. Pour l'analyse acoustique on ne s'intéresse qu'à la partie stable de la voyelle, mais la validation perceptive de ces voyelles est faite par écoute de l'ensemble du segment CVC. Certains auteurs ont obtenu des résultats d'identification nettement meilleurs avec un contexte CVC plutôt que hors contexte [Strange et coll. 1976]. Ces résultats ont été infirmés par la suite [Kahn 1978, Macchi 1980], avec une meilleure sélection des locuteurs et des auditeurs (pour s'assurer qu'ils ont la même origine linguistique) et un contrôle plus strict des conditions d'élocution.

L'importance de la tâche assignée aux auditeurs a été mise en évidence par divers auteurs [Assman et coll. 1982] : le comportement des auditeurs diffère selon qu'ils doivent choisir une catégorie phonétique, choisir un mot comprenant la même voyelle, répéter le son qu'ils ont perçu.

	/A/	/i/	/u/	/e/	/ɛ/	/y/	/ø/	/oe/	/o/	/ɛ̃/	/ɑ̃/	/ɔ̃/	??	tot	err	%
/A/	58									1	1			60	2	3
/i/		54		2		3							1	60	6	10
/u/			58						1				1	60	2	3
/e/				58									2	60	2	3
/ɛ/				1	55			1					3	60	5	8
/y/				1		59								60	1	2
/ø/							58	2						60	2	3
/oe/							13	47						60	13	22
/o/			4						55				1	60	5	8
/ɛ̃/								2		54	1		3	60	6	10
/ɑ̃/	1									1	52	2	4	60	8	13
/ɔ̃/			2				1		1		3	46	7	60	14	23
tot	59	54	64	62	55	62	72	52	57	56	57	48	22	720	66	
dif	-1	-6	4	2	-5	2	12	-8	-3	-4	-3	-12				

Tableau 2 : confusions phonétiques faites majoritairement par les auditeurs, tous locuteurs et tous styles confondus. La ligne "tot" représente le total de chaque colonne. La ligne "dif" représente le pouvoir attracteur, c'est-à-dire la différence, pour chaque voyelle, entre le total en colonne ("voyelle perçue") et le total en ligne ("voyelle requise").

L'ordre de présentation des stimuli peut être "bloqué" ("blocked"), tous les signaux d'une même série provenant d'un même locuteur, ou "mixte" ("mixed") : dans ce cas les signaux provenant de divers locuteurs sont présentés aléatoirement ([Strange et coll. 1976], [Verbrugge et al. 1976]), si bien que les auditeurs n'ont pas la possibilité d'utiliser des informations d'adaptation au nouveau locuteur. Le problème avait été déjà évoqué par Ainsworth [Ainsworth 1975] qui a introduit les termes de "perception intrinsèque" et "perception extrinsèque".

Le tableau 2 résume les confusions phonétiques pour l'ensemble des locuteurs, tous styles confondus. Le nombre total de confusions et de non-décisions est de 66 (soit 9.2%). Les voyelles les plus mal identifiées sont /œ̃/ et /oe/ , qui donnent lieu respectivement à 23% et 22% d'erreur. Les voyelles les mieux identifiées sont /A/, /u/, /e/, /y/, avec moins de 3% d'erreur.

On notera que /oe/ est souvent identifiée comme /ø/, beaucoup plus que /ø/ comme /oe/. Il s'agit d'une particularité régionale de nos locuteurs, dont plusieurs ont pour origine la région Sud-Est de la France. Sur les 13 erreurs /oe/ -> /ø/, 10 proviennent de locuteurs d'origine grenobloise, et 6 parmi ces 10 sont imputables à un même locuteur. Par contre les erreurs observées sur /œ̃/ ne bénéficient pas systématiquement à une autre voyelle, et l'on observe alors un maximum de non-reconnaissances.

On a reporté au bas du tableau 2 deux lignes représentant respectivement le total observé pour chaque colonne, c'est-à-dire le score de chaque voyelle en tant que "voyelle perçue", et la différence par rapport au total en ligne (qui représente le score en tant que "voyelle requise"). Cette différence, ou "pouvoir attracteur", montre que certaines voyelles (que nous qualifierons de "fortes") attirent les suffrages qui devraient aller à d'autres, et que, réciproquement, certaines (que nous qualifierons de "faibles") sont plus fragiles. Le glissement /oe/->/ø/ est ainsi bien mis en évidence : pour les locuteurs de notre corpus il est manifeste que /ø/ est une voyelle forte, et /oe/ une voyelle faible. On voit aussi que /u/ est plutôt une voyelle forte, que /œ̃/ est très faible, que /i/, /ɛ/ et /ɛ̃/ sont plutôt faibles.

Nos résultats montrent une grande différence entre les voyelles orales et les voyelles nasales. En effet, nous avons pour les 9 voyelles orales un total de 38 erreurs, soit en moyenne 7% par voyelle, alors que les 3 voyelles nasales produisent en tout 28 erreurs, soit en moyenne 16% par voyelle. Si l'on considère le pouvoir attracteur, la différence entre orales et nasales apparaît encore plus nettement : en moyenne celui-ci est quasi-nul pour les orales (-0.1), alors qu'il est notablement négatif (-6.3) pour les nasales. Cela signifie que les nasales ont tendance à être prises pour des orales, mais pas l'inverse.

La répartition du nombre d'erreurs selon le locuteur montre que le taux d'erreur varie entre 0 et 31% (tableau 1). Certaines erreurs paraissent liées à un parler régional et sont plutôt de nature linguistique (cas de la confusion /oe/->/ø/ chez CB, par exemple), d'autres semblent caractériser individuellement le locuteur et sont plutôt de nature diagnostique (cas de la mauvaise prononciation de /œ̃/ chez ML, de /ɑ̃/ chez SB et de la confusion /i/->/y/ chez MF).

Nos résultats pour l'identification phonétique peuvent être comparés à ceux de Assman et coll. [Assman et coll. 1982], avec toutes les réserves tenant à la différence de langue et aux différences de protocole expérimental. Pour 10 voyelles de l'anglais canadien d'Edmonton, provenant de 10 locuteurs (5 F et 5 H), prononcées isolément, hors contexte consonantique, présentées en ordre aléatoire, ces auteurs ont défini des taux d'erreur de 11% (lorsque les auditeurs répondent au moyen d'un mot-clé /hVd/), et 9% (lorsqu'ils répondent avec un mot-clé /pVp/). Avec 9.2% nos résultats sont très voisins, mais cette coïncidence ne doit pas masquer les différences évoquées ci-dessus. Dans les deux cas on a constaté de forts écarts d'une voyelle à l'autre (de 1 à 43% chez Assman et coll. 1982); nous avons, en ce qui nous concerne, minimisé cet effet en excluant les deux variantes du /A/, ainsi que /ɔ/ et /œ̃/.

V - DESCRIPTEUR "GENRE DU LOCUTEUR"

La perception de ce descripteur diagnostique est rarement étudiée en tant que telle, mais le problème peut apparaître dans les conditions usuelles de la communication, par exemple au téléphone.

Notre but ici n'est pas de déterminer quels paramètres physiques sont à l'origine de la perception d'une voix masculine ou féminine, mais plutôt de mettre en évidence d'éventuelles relations entre ce descripteur et les descripteurs "locuteur" et "identité de la voyelle".

Le tableau 1 montre que, à deux exceptions près, seul le locuteur JB (féminin) est concerné. Sur les 108 segments produits par ce locuteur, 27 ont été perceptivement attribués à une voix masculine ou déclarés ambigus. Si l'on examine les erreurs en fonction de la voyelle requise, tous styles confondus, il apparaît que les erreurs se produisent surtout pour /y/, /u/, /ø/ et /œ̃/, c'est-à dire lorsque la voyelle est arrondie.

Dans notre corpus nous n'avons observé qu'un petit nombre d'erreurs sur le genre du locuteur. Il ne faudrait pas en déduire que le problème de percevoir

ce descripteur est secondaire. Il est probable que si nous avions utilisé comme locuteurs des adolescents ou de jeunes enfants le problème se serait posé fréquemment.

VI - DESCRIPTEUR "STYLE DE VOIX"

La relation entre "style requis" et "style perçu" fait intervenir des considérations complexes de la part du locuteur (compréhension de la consigne, ajustement de la force de voix au niveau requis pour la voyelle requise, contrôle proprioceptif ou auditif), et de la part de l'auditeur: ce dernier doit évaluer l'intention du locuteur de parler plus ou moins fort, mais son jugement est influencé par le niveau sonore, dans l'absolu, du son qu'il écoute.

On ne trouve que peu de travaux concernant l'effort de parole ou le style de voix (Schulman 1985, Traunmüller 1985, Granström and Nord 1991). Ces travaux sont souvent en rapport avec les mécanismes de production de la parole. Du point de vue perceptif, on sait que le seuil de reconnaissance varie selon les voyelles [Wajskop 1971] : à intensité physique égale, les voyelles compactes, dont l'énergie est concentrée dans le centre du spectre (autour de 1000 Hz), sont perçues avec une plus grande intensité subjective que les voyelles diffuses, dont l'énergie est plutôt située vers le grave ou vers l'aigu. Les études menées sur ce sujet sont de nature psycho-acoustique, avec des stimuli calibrés qui n'ont qu'un rapport lointain avec les matériaux utilisés dans la présente étude.

La gamme de variation des styles de parole dans laquelle nous nous sommes situés est celle de la vie courante. Plus précisément nous avons voulu explorer l'intervalle de niveau sonore dans lequel on adapte l'effort vocal à la situation de communication de manière pratiquement inconsciente. Les premières mesures acoustiques montrent que d'un style à l'autre la variation moyenne est de 6 à 10 dB, soit moins de 20 dB entre extrêmes, ce qui est très peu par rapport à la dynamique possible de la voix.

Chaque ligne du tableau 3 représente la ventilation des 240 segments émis selon un même "style requis", dans les trois catégories de "style perçu" que nous avons déterminées plus haut.

	P	N	L	tot
P	162	70	8	240
N	66	138	36	240
L	0	27	213	240

Tableau 3 : style perçu, en fonction du style requis, tous locuteurs confondus

La somme des valeurs de la partie haute (au-dessus de la diagonale principale) représente le nombre de cas où, selon les auditeurs, la voix a été produite plus fortement que ne l'indiquait la consigne. Inversement la partie basse (au dessous de la diagonale principale) représente les cas de production moins forte que la consigne. Nous appellerons pmf ("plus fort/moins fort") le rapport des deux parties, et nous l'exprimerons sous forme logarithmique (10*log(pmf)) par symétrie. Ici ce rapport est de 1.22, soit +0.9 sous forme logarithmique. Cette valeur faible indique que les erreurs de style sont équilibrées dans l'ensemble.

On a reporté dans la figure 1 les deux indices caractérisant les erreurs faites sur le style, en fonction de l'identité de la voyelle requise. Il apparaît un quadrilatère /u/ /ɑ̃/ /ɛ̃/ /Ø/, à l'intérieur duquel les autres voyelles se regroupent de manière cohérente. Selon les ordonnées (rapport pmf) on voit que les voyelles /y/, /u/ et /Ø/ (arrondies, produites avec une protrusion des lèvres) tendent à induire une voix moins forte que ce qui est requis, alors que les nasales /ɑ̃/, /ɔ̃/ et /ɛ̃/ tendent à induire une voix plus forte. Selon les abscisses (taux d'erreur) on voit que /y/, /A/ et /ɑ̃/ donnent lieu à peu d'erreurs de style, alors que /ɛ̃/, /e/, /i/ et /Ø/ produisent plus d'erreurs, mais celles-ci sont peu caractéristiques.

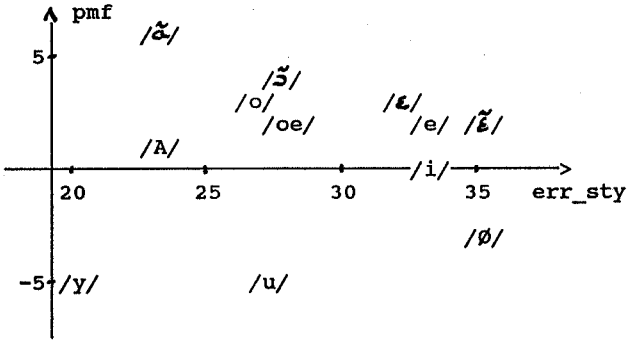


Figure 1 - Nature des erreurs de style (rapport pmf) en fonction du taux de ces erreurs, pour les diverses voyelles

Le taux d'erreur sur le style et le rapport pmf sont indiqués dans le tableau 1 pour chaque locuteur. On constate de grandes différences individuelles. Le rapport pmf indique dans quelle mesure, lorsqu'il a fait des erreurs de style, le locuteur a parlé moins fort ou plus fort que ce qui était requis.

Le locuteur JB (féminin) est le seul pour lequel on ait obtenu des évaluations de "genre perçu" masculin. Dans le tableau 4 est reporté le nombre de réponses obtenu pour chaque combinaison des descripteurs "genre perçu" et "style perçu".

	P	N	L
H	10	5	2
F	26	19	36
?	3	4	3

Tableau 4 : Répartition des 108 évaluations relatives au locuteur JB, sur les descripteurs "style perçu" et "genre perçu".

Il est clair que la confusion avec une voix masculine se produit lorsque la locutrice parle à voix moyenne ou faible. La répartition des erreurs selon la voyelle requise montre que les erreurs sur le genre du locuteur se produisent essentiellement sur les voyelles arrondies, dont on a vu plus haut qu'elles étaient moins sonores que les autres.

On peut remarquer que la locutrice JB, pratiquement seule à produire des erreurs de genre, produit très peu d'erreurs phonétiques, et très peu d'erreurs de style, et que ces erreurs sont d'autant plus importantes que la force de voix est moyenne ou faible. On peut penser que c'est justement parce qu'elle marque bien les divers styles qu'elle est amenée à exagérer la consigne, et à adopter une voix très faible quand une voix faible suffirait. Elle est amenée ainsi à baisser exagérément son fondamental, ce qui contribue à donner un caractère masculin à sa voix.

Ainsi se dessine pour ce locuteur une certaine stratégie d'élocution. Cet exemple montre bien les relations entre les trois descripteurs. Il suggère aussi que la connaissance ou la reconnaissance des descripteurs diagnostiques pourrait être mise à profit dans un système automatique pour mieux reconnaître la voyelle prononcée, ou du moins pour prédire la nature des erreurs possibles dans des circonstances données: ici la connaissance de la valeur "voix faible" du descripteur "style" permet d'augurer des erreurs de type F->H sur le genre du locuteur et des confusions sur les voyelles arrondies.

VII - CONCLUSION

Cette étude a un résultat tangible, sous la forme de deux fichiers, l'un comportant les signaux vocaliques, l'autre la description symbolique de ces éléments. Cette description, qui se veut aussi complète que possible, porte sur l'identité phonétique de la voyelle, sur le genre du locuteur, et sur le style de voix; elle reflète d'une part la consigne donnée au locuteur, d'autre part l'évaluation qui en a été faite par un groupe d'auditeurs. Ainsi ces données peuvent être utilisées en connaissance de cause dans des études de traitement automatique de la parole.

A partir des descriptions symboliques on a pu trouver des résultats perceptifs voisins de ceux obtenus dans une autre langue en ce qui concerne l'identité phonétique. On a pu aussi mettre en évidence la capacité du canal vocal à transmettre, avec de simples voyelles isolées, des informations non-linguistiques comme la force de voix et le genre du locuteur. De plus on a montré que les erreurs phonétiques, les erreurs sur le genre du locuteur et les erreurs sur la force de voix sont liées entre elles, et varient d'un locuteur à l'autre. On a ainsi justifié l'intérêt d'une approche qui prend en compte, ensemble, tous les aspects perceptifs de la parole et de la voix, au lieu de chercher dans le signal de stricts invariants linguistiques.

VII - REFERENCES

Ainsworth, W. (1975): "Intrinsic and extrinsic factors in vowel judgments", in Auditory analysis and perception of speech, ed. by G.Fant and M.Tatham, Academic, London, 103-113.

Assman, P.F., Nearey, T.M. and Hogan, J.T. (1982): "Vowel identification: orthographic, perceptual and acoustic aspects", J.Acoust.Soc.Am 71 (4), 975-989.

Granström, B. and Nord, L. (1991): "Neglected dimensions in speech synthesis", ESCA workshop on The phonetics and phonology of speaking styles, Barcelona.

Kahn, D. (1978): "On the identifiability of isolated vowels", UCLA Working Pap. Phon. 41, 26-31.

Liénard, J.S. (1990): "Perception, data variability and inductive inference", Cognitiva, AFCET, Madrid.

Liénard, J.S. et Di Benedetto, M.G. (1992): "CORENC : un corpus de voyelles françaises isolées et son évaluation perceptive selon divers descripteurs", Rapport Interne LIMSI.

Macchi, M.J. (1980): "Identification of vowels spoken in isolation vs vowels spoken in consonantal context", J.Acoust.Soc.Am. 68 (6), 1636-1642.

Peterson, G.E. and Barney, H.L. (1952): "Control methods used in a study of the vowels", J.Acoust.Soc.Am. 24 (2), 175-184.

Schulman, R. (1985): "Articulatory targeting and perceptual constancy of loud speech", Séminaire franco-suédois, ICP, Grenoble.

Traunmüller,H. (1985): "The role of the fundamental and the higher formants in the perception of speaker size, vocal effort, and vowel openness", Séminaire franco-suédois, ICP, Grenoble.

Verbrugge, R.R., Strange W., Shankweiler, D.P. and Edman T.R. (1976): "What information enables a listener to map a talker's vowel space ?", J.Acoust.Soc.Am. 60 (1), 198-212.

Wajskop, M. (1971): "Seuils de reconnaissance de voyelles isolées", Revue d'Acoustique 13, 20-22.

19es J.E.P. - Bruxelles, 19 au 22 mai 1992.

IMPORTANCE DES DIFFERENTS FACTEURS DE VARIABILITE INTERNE AU GROUPES DE CONSONNES.

MEUNIER Christine

INSTITUT DE PHONETIQUE, UNIVERSITE DE PROVENCE
29, Av. Robert Schuman, 13621 AIX-EN-PROVENCE

Résumé

Dans le cadre d'une étude générale sur la variabilité acoustique des groupes de consonnes, nous nous proposons ici d'étudier le problème particulier de la variabilité interne au groupe de consonnes (GC), autrement dit, la variabilité qui est due aux unités phonétiques constituantes du GC (mode d'articulation, mode de phonation, lieu d'articulation). Afin d'analyser ce second type de variabilité, nous avons utilisé les corpus ACC01 à ACC05 de la Base de Données des Sons du Français (BDSONS). Nous nous intéressons ici particulièrement aux variations spectrales des consonnes du groupe. L'importance de cette variation est analysée et évaluée en quantité (pourcentage de variation sur la totalité de la durée de la consonne), et en qualité (type de variation subie: déplacement de l'énergie vers le haut ou vers le bas du spectre).

1-INTRODUCTION

Les groupes de consonnes sont rarement appréhendés dans leur ensemble. Lorsqu'ils le sont, l'approche utilisée est distributionelle (Rossi, 1968, Aubergé et al., 88) ou articulatoire (Rochette, 1973). Les études acoustiques des GC concernent la plupart du temps des groupes précis comme les groupes d'occlusives (Marchal, 1985), ou, très fréquemment les groupes occlusives+liquides. Toutefois, les GC ne sont jamais étudiés acoustiquement de façon globale... et on le comprend aisément! Etudier acoustiquement l'ensemble des GC, n'est-ce pas supposer qu'il existe une homogénéité dans la réalisation acoustique des consonnes? Cette position est difficilement tenable: on ne peut, au sens strict des indices acoustiques, considérer qu'il existe un ensemble de consonnes différent d'un ensemble de voyelles; en effet, il n'y a acoustiquement pas plus de différence entre [l] et [a], qu'entre [l] et [p].

Ainsi, si l'on ne peut trouver d'emblée de statut acoustique à l'ensemble des consonnes, et donc au groupes de consonnes, on cherchera plutôt leur statut d'unité dans leur organisation syntagmatique au sein de la syllabe. Hjelmslev (1963) considère que les voyelles sont "des grandeurs établissant par elles-mêmes une syllabe" alors que les consonnes sont "des grandeurs n'établissant pas par elles-mêmes une syllabe". Autrement dit, il caractérise les voyelles comme des éléments "centraux", tandis que les consonnes seraient des éléments "périphériques" dans la syllabe.

On peut à partir de ce postulat syntagmatique revenir à une approche acoustique de l'ensemble des GC. En effet, il est possible d'envisager que ce statut facultatif de la consonne dans l'unité syllabique soit repérable sous la forme d'indices acoustiques (on pense ici à une plus forte variabilité ou une instabilité temporelle ou encore une importante coarticulation). C'est dans cette perspective que nous avons entrepris une étude de la variabilité acoustique de l'ensemble des GC du français. Nous pensons en effet que les indices observés dans les phénomènes de variation au sein des GC seront riches d'informations permettant de cerner les particularités d'un "espace consonantique".

2-VARIABILITE ET VARIATION.

La variabilité est un domaine très complexe. Elle l'est d'autant plus lorsqu'il s'agit de l'évaluer sur des combinaisons d'unités phonétiques.

Distinguons tout d'abord causes et conséquences: certains facteurs provoquent des changements à l'intérieur du groupe de consonnes (le type de GC, la place dans le mot, etc), ce sont les *causes* que nous nommerons désormais *facteurs de variabilité;* puis il y a l'effet lui-même qui est produit par les facteurs de variabilité sur le GC (dévoisement, allongement, bémolisation, etc), ce sont les conséquences que nous nommerons désormais *variations.*

Concernant les facteurs de variabilité, nous concevons deux types de variables à prendre en compte: les *variables internes,* d'une part, qui concernent les éléments constitutifs des GC (mode d'articulation, lieu d'articulation et mode de phonation des consonnes du groupe) et les *variables externes,* d'autre part, qui dépendent de facteurs non constitutifs des GC mais qui

sont susceptibles d'entraîner des variations acoustiques à l'intérieur même du groupe (place dans le mot, place dans la syllabe, rôle de l'accent, type de prononciation, débit, voyelle adjacente, etc).

Dans cette communication, nous nous intéressons précisément aux conséquences spectrales des variables internes. Nous évaluons, par conséquent, l'impact du mode et du lieu d'articulation ainsi que du mode de phonation d'une consonne du GC sur la répartition spectrale de l'énergie de l'autre consonne.

3-HYPOTHESES ET QUESTIONS

L'ensemble de cette étude est fondée sur l'hypothèse suivante: La variabilité des caractéristiques acoustiques des lieux d'articulation de C1 (par exemple /s/: aigü; /f/: grave) serait susceptible d'attirer la répartition spectrale de l'énergie de C2 vers les hautes ou les basses fréquences.

Cette hypothèse, une fois admise, laisse quelques interrogations en suspens:

- Cette variation de la répartition spectrale de C2 est-elle également dépendante du mode d'articulation et de phonation de C1?
- On suppose une variation de C1 vers C2. L'inverse est-il possible et dans quelle mesure?
- Quelle est cette variation? Autrement dit, comment la décrire: est-ce l'ensemble de la consonne qui subit la variation ou seuleument une partie? La variation porte-elle- sur les hautes ou sur les basses fréquences?

Nous tenterons ici de vérifier l'hypothèse de base et de répondre aux questions qui en découlent. Nous présentons ci-dessous un ensemble méthodologique conçu dans l'optique de répondre précisément à ces interrogations.

4-METHODOLOGIE.

a-Classification en consonnes et groupes de consonnes:

De toutes les possibilités de classer les consonnes, une classification en mode d'articulation est celle qui convient le mieux à une analyse sur les groupes de consonnes. Nous regroupons ainsi l'ensemble des consonnes du français en trois classes (Autesserre et Rossi, 1987):

- Les occlusives (OCC): voisées /b/ /d/ /g/ et non voisées /p/ /t/ /k/.
- Les fricatives (FRI): voisées /v/ /z/ /Z/ et non voisées /f/ /s/ /S/.
- Les consonnes vocaliques (C.VOC): nasales /m/ /n/, liquides: /l/, /r/, glissantes: /j/, /w/, /y/.

De ce classement en consonnes, nous déduisons deux types de GC (Meunier, 1990):

- Les groupes consonantiques homogènes (GC_{ho}) où les deux consonnes appartiennent à la même classe de consonnes:

$GC_{ho}1$ = OCC + OCC
$GC_{ho}2$ = FRI + FRI
$GC_{ho}3$ = C.VOC + C.VOC

- Les groupes consonantiques hétérogènes ($GC_{hé}$) dans lesquels les deux consonnes appartiennent à deux classes de consonnes différentes:

$GC_{hé}1$ = OCC + FRI
$GC_{hé}2$ = FRI + C.VOC
$GC_{hé}3$ = OCC + C.VOC

b-Matériel linguistique

Nous avons utilisé les corpus ACC01 à ACC05 de la Base de Données des Sons du Français (BDSONS, Carré et al., 1987). Il s'agit de cinq corpus de 41 mots (lus isolément) contenant chacun un groupe de consonnes et parmi lesquels nous avons retenu ceux qui sont pertinents pour notre étude. Pour sélectionner l'ensemble de notre corpus, nous avons utilisé un gestionnaire de bases de données (GERSONS, développé à l'I.C.P, Grenoble) et le CDROM BDSONS_2 dans lequel sont stockés les corpus ACC01 à ACC05. Ces corpus sont lus par 12 locuteurs, 6 femmes et 6 hommes. Afin de neutraliser l'effet éventuel d'autres variables, nous ne comparons que des paires minimales différant par l'un des trois facteurs de variabilité interne cités ci-dessus: /bru/-/dru/-/gru/ (lieu d'articulation), /klu/-/flu/ (mode d'articulation), /bri/-pri/ (mode de phonation). La voyelle reste évidement toujours la même dans un groupe de comparaison donné.

c-Analyse acoustique

Nous avons effectué des analyses spectrales sur les consonnes des mots du corpus à l'aide du logiciel d'étiquetage PTS (développé à l'I.C.P, Grenoble). La mesure spectrale dépend évidemment de la consonne étudiée. En ce qui concerne les consonnes vocaliques, on mesure le spectre de F1 et F2 au centre et au 3/4 de la consonne (ceci nous permet d'évaluer la variation sur l'échelle spectrale ainsi que sur l'échelle temporelle). Pour les fricatives, on mesure la zone de bruit dont l'énergie est maximale au centre et au 3/4 de la consonne. Enfin, on mesure la répartition de l'énergie sur l'explosion de l'occlusive.

5-RESULTATS ET INTERPRETATIONS

a-Position C1

Dans le corpus que nous avons pu rassembler dans BDSONS, la première consonne du groupe était soit une consonne vocalique (groupes du type Ho3, "lui"), soit une fricative (groupes du type Hé2, "flou"), soit une occlusive (groupes du type Hé3, "bras").

Consonnes vocaliques

	cent-a	cent-b	3/4-a	3/4-b
lui	199-24	1774-10	194-22	1796-8
louis	223-29	1595-31	226-30	1456-30

Tableau 1: moyennes (en Herz) et coefficients de variation des valeurs de /l/."a" représente les valeurs de F1 (basses fréquences) et "b" les valeurs de F2; "cent" et 3/4 signifient respectivement que les mesures ont été prises au centre et à la fin de la consonne

	voc1a	voc1b	bat-a	bat-b	voc2a	voc2b
ruis	277 49	1023 33	167 34	825 40	277 43	1049 28
rouis	328 34	751 30	265 57	1131 87	190 50	823 68

Tableau 2: moyennes (en Herz) et coefficients de variation des valeurs de /r/."a" représente les valeurs de F1 (basses fréquences) et "b" les valeurs de F2; "voc1", "bat", et "voc 2" signifient respectivement que les mesures ont été prises au centre de la première partie vocalique, au centre du battement et au centre de la deuxième partie vocalique.

Les consonnes vocaliques en position de C1 subissent notablement l'influence de C2. Cette influence se manifeste de la façon suivante: /w/ tend à bémoliser les valeurs de C1, alors que /y/ attire les valeurs de C1 vers les hautes fréquences. Le /l/ subit une variation plus importante au 3/4 qu'au centre (tableau 1); tandis que /r/ varie essentiellement pour F2. Il est intéressant de constater que le /r/, malgré son battement central, subit la variation de C2 dans sa totalité.

Fricatives

La seule fricative que nous permet d'étudier notre corpus, dans cette position, est /f/. C'est malheureusement celle dont la répartition spectrale est la plus difficile à délimiter. Il nous faudra donc en tenir compte dans l'interprétation de nos résultats.

	centre	3/4
foi	2183-44	1159-33
flash	1998-31	2184-39
frappe	2315-39	2232-44
flic	1988-40	1876-19
fripe	2250-35	1525-29
flou	2044-54	1543-45
frup	1887-66	1412-50

Tableau 3: moyennes (en Herz) et coefficients de variation des valeurs de /f/. "centre" et 3/4 signifient respectivement que les mesures ont été prises au centre et à la fin de la consonne

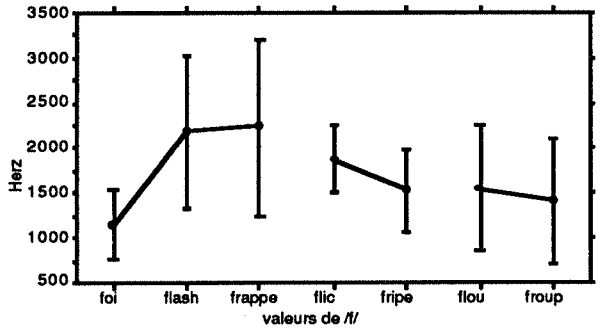


Figure 1: moyennes et écarts types des valeurs de /f/ mesurées au 1/4 de la consonne. En abcisse les fréquences en Herz, en ordonnée les contextes de réalisation du /f/. Nous avons relié par un trait continu les contextes vocaliques identiques.

On constate ici que le /f/ ne semble pas subir significativement l'influence de /r/ ou de /l/. Toutefois, la glissante /w/ semble attirer notablement la zone de bruit de /f/ vers les basses fréquences mais seulement au 3/4 de la consonne.

Les occlusives

Nous n'avons pu étudier spécifiquement les occlusives. La mesure spectrale du burst s'avère trop aléatoire: l'explosion de /b/ ou de /p/ est souvent non identifiable et la variation de la fréquence du burst ne s'est pas avérée significative. Nous verrons cependant que l'étude des variations de C2 nous informe sur la nature de l'occlusive lorsqu'elle est en position de C1.

b-Position C2

Notre corpus n'est constitué que de consonnes vocaliques en position C2. Cela n'a rien d'étonnant, car dans ce que l'on appelle groupes consonantiques (ou "clusters", Pulgram, 1965), les structures les plus fréquentes sont celles qui correspondent à nos classes Hé2 et Hé3.

Nous évaluons ainsi la variation subie par les consonnes vocaliques en position C2, et précédées respectivement d'une autre consonne vocalique (Ho3), d'une fricative (Hé2) et d'une occlusive (Hé3).

Les consonnes vocaliques dans Ho3

Pour notre corpus, les consonnes vocaliques en position C2 dans les groupes Ho3 sont des glissantes:

	1/4-a	1/4-b	cent-a	cent-b
louis	277-26	1196-11	225-29	719-19
rouis	235-34	646-21	256-26	732-18
lui	244-27	1975-12	214-20	1903-7
ruis	295-25	1402-22	246-25	1530-27
nuit	212-26	2217-30	221-30	1840-27
loi	314-29	1124-13	256-36	598-23
roi	248-54	597-25	229-53	622-21
noix	255-39	1167-12	217-52	602-24
moi	237-48	685-29	288-50	698-23

Tableau 4: moyennes (en Herz) et coefficients de variation des valeurs des glissantes /w/ et /y/."a" représente les valeurs de F1 (basses fréquences) et "b" les valeurs de F2; "cent" et 3/4 signifient respectivement que les mesures ont été prises au centre et à la fin de la consonne.

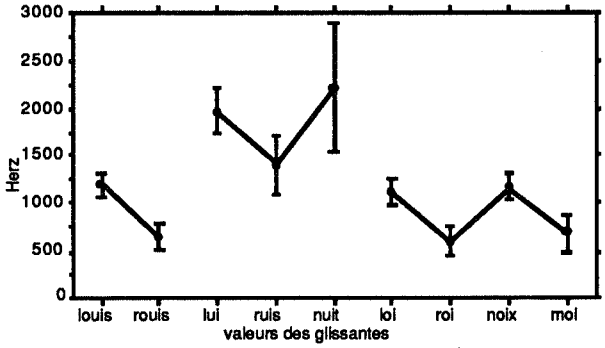


Figure 2: moyennes et écarts types des valeurs du F2 des glissantes /w/ et /y/ mesurées au 1/4 de la consonne. En abcisse les fréquences en Herz, en ordonnée les contextes de réalisation. Nous avons relié par un trait continu les mêmes glissantes suivies de contextes vocaliques identiques.

La variation subie par C2 dans les groupes Ho3 est importante, significative et régulière (figure 2). On

peut la résumer ainsi: /l/ et /n/ augmentent les valeurs du F2 des glissantes, tandis que /r/ et /m/ diminuent ces valeurs. Ces variations n'affectent que le F2 au 1/4 de la glissante (tableau 4). Notons que pour la figure ci-dessus, les écarts types sont relativement faibles.

Nous constatons sur la figure 2 que la variation apportée par le changement de contexte /l/-/r/ se fait indépendamment de la voyelle adjacente (qui influence également l'ensemble du GC, Meunier, 1992) et de la glissante étudiée (ici, /w/, /y/).

Les consonnes vocaliques dans Hé2

Dans notre corpus, les consonnes vocaliques en position C2 dans Hé2 sont /l/ et la glissante /w/.

	1/4-a	1/4-b	cent-a	cent-b
flash	346-20	1674-14	391-18	1761-23
slash	299-25	1937-28	319-28	2035-32
flic	236-37	2037-22	228-28	2179-34
slip	219-46	2340-20	199-23	2303-23
flou	219-38	1386-17	239-20	1451-12
sloop	212-23	1719-13	199-37	1661-14

Tableau 5: moyennes (en Herz) et coefficients de variation des valeurs de /l/."a" représente les valeurs de F1 (basses fréquences) et "b" les valeurs de F2; "cent" et 3/4 signifient respectivement que les mesures ont été prises au centre et à la fin de la consonne

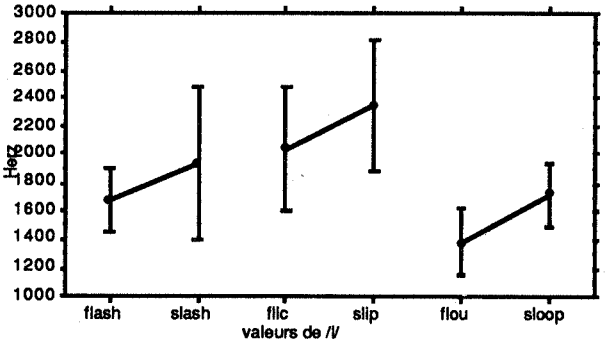


Figure 3: moyennes et écarts types des valeurs du F2 de /l/ mesurées au 1/4 de la consonne. En abcisse les fréquences en Herz, en ordonnée les contextes de réalisation. Nous avons relié par un trait continu les contextes vocaliques identiques.

En ce qui concerne les variations subies par /l/, on observe une opposition grave/aigü occasionnée respectivement par /f/ et /s/. Cette variation est observable au 1/4 de la consonne.(figure 3) Elle est toujours présente au centre mais tend à s'atténuer (tableau 5). L'influence de C1 est présente indépendamment de celle de la voyelle adjacente.

	1/4-a	1/4-b	cent-a	cent-b
foi	246-43	581-26	384-24	780-18
soi	438-44	1375-13	221-40	644-28
choix	460-48	1250-13	228-37	630-27
voix	230-41	656-21	317-37	700-23
joie	241-31	1283-19	251-35	641-25

Tableau 6: moyennes (en Herz) et coefficients de variation des valeurs de /w/."a" représente les valeurs de F1 (basses fréquences) et "b" les valeurs de F2; "cent" et 3/4 signifient respectivement que les mesures ont été prises au centre et à la fin de la consonne

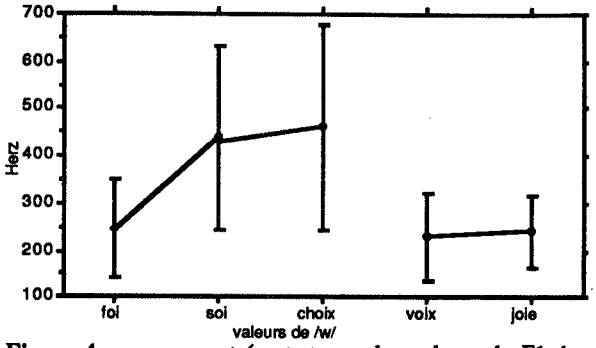


Figure 4: moyennes et écarts types des valeurs du F1 de /w/ mesurées au 1/4 de la consonne. En abcisse les fréquences en Herz, en ordonnée les contextes de réalisation. Nous avons relié par un trait continu les contextes vocaliques identiques en fonction du mode de phonation de C1.

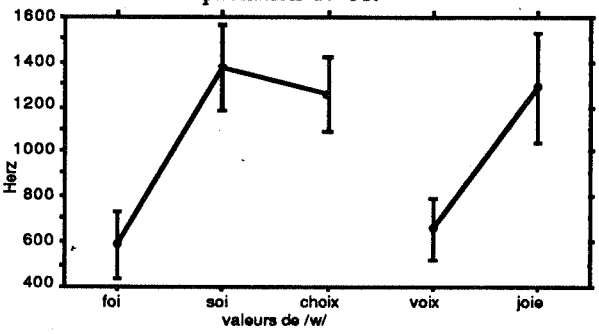


Figure 5: moyennes et écarts types des valeurs du F2 de /w/ mesurées au 1/4 de la consonne. En abcisse les fréquences en Herz, en ordonnée les contextes de réalisation. Nous avons relié par un trait continu les contextes vocaliques identiques en fonction du mode de phonation de C1.

Avec la glissante /w/, nous avons l'occasion d'évaluer l'influence des trois lieux d'articulation des fricatives. Une fois de plus, c'est au 1/4 de la consonne que la variation est observable. Pour F1 (figure 4), le /f/ entraîne les valeurs de /w/ vers les basses fréquences, ce qui n'est plus vrai lorsque l'on fait varier le mode d'articulation: il n'y a pas de différence significative entre les valeurs de /w/ dans "voix" et celles de /w/ dans "joie". Par contre, /f/ et /v/ font varier /w/ de la même façon si l'on observe F2 (figure 5).

La consonne vocalique dans Hé3

Dans notre corpus, les consonnes vocaliques en position C2 dans Hé3 sont /l/, /r/ et les glissantes /w/./y/.

	1/4-a	1/4-b	cent-a	cent-b
pli	261-38	2031-20	214-20	2085-22
clic	234-30	2682-19	248-23	2595-23
plouc	246-33	1366-11	249-38	1425-14
clou	268-51	1521-17	212-24	1483-14
blague	339-28	1780-13	369-27	1718-13
glas	261-28	2563-28	290-16	1950-33

Tableau 7: moyennes (en Herz) et coefficients de variation des valeurs de /l/."a" représente les valeurs de F1 (basses fréquences) et "b" les valeurs de F2; "cent" et 3/4 signifient respectivement que les mesures ont été prises au centre et à la fin de la consonne.

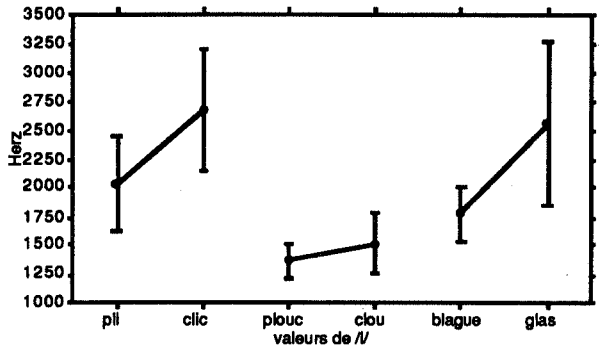


Figure 6: moyennes et écarts types des valeurs du F2 de / l / mesurées au 1/4 de la consonne. En abcisse les fréquences en Herz, en ordonnée les contextes de réalisation. Nous avons relié par un trait continu les contextes vocaliques identiques en fonction du mode de phonation de C1.

Comme dans les groupes Hé2 analysés précédemment, on constate que le /l/ subit l'influence du lieu d'articulation de C1. Ici, on observe que les labiales /p/ et /b/ attirent les valeurs du F2 de /l/ vers les basses fréquences, alors que les vélaires /k/ et /g/ attirent F2 vers les hautes fréquences (figure 6). On notera cependant que cette variation est quasiment inexistante devant la voyelle /u/. Nous donnerons ci-après une interprétation de ce phénomène. La variation observée ne semble pas dépendante du mode de phonation de C1; elle est toujours présente au centre du /l/, mais de façon plus atténuée (tableau 7).

	1/4-a	1/4-b	cent-a	cent-b
toi	360-39	1384-19	217-54	624-30
quoi	327-42	776-28	317-35	697-21
bois	239-38	773-13	342-22	726-19
doigt	248-39	1382-22	277-19	731-12
buis	187-22	1952-13	228-33	1932-10
duis	180-31	2316-19	194-19	1854-10

Tableau 8: moyennes (en Herz) et coefficients de variation des valeurs des glissantes /w/ et /y/."a" représente les valeurs de F1 (basses fréquences) et "b" les valeurs de F2; "cent" et 3/4 signifient respectivement que les mesures ont été prises au centre et à la fin de la consonne

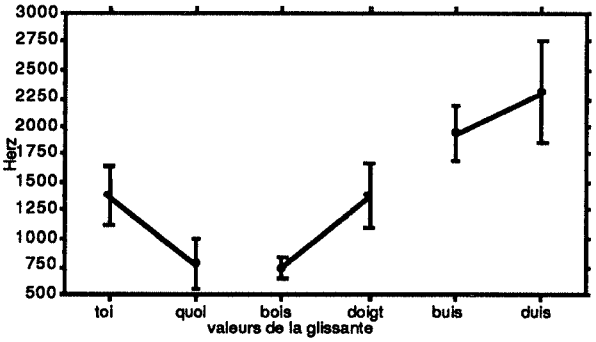


Figure 7: moyennes et écarts types des valeurs du F2 de /w/ et /y/ mesurées au 1/4 de la consonne. En abcisse les fréquences en Herz, en ordonnée les contextes de réalisation. Nous avons relié par un trait continu les contextes vocaliques identiques en fonction du mode de phonation de C1.

Les glissantes subissent également l'influence de C1, toujours au 1/4 de leur durée et pour F2. La variation subie est indépendante de la voyelle adjacente et du mode de phonation de C1. Les dentales /t/ et /d/ attirent le F2 des glissantes vers les hautes fréquences. Ce corpus ne nous permet pas de distinguer les rôles des labiales et des vélaires.

	voc1a	voc1b	bat-a	bat-b	voc2a	voc2b
bras	279 65	932 28	239 62	1030 28	317 66	1208 41
drap	440 32	1342 22	131 57	1062 25	304 71	1127 11
gras	292 55	1469 8	234 60	1201 23	361 71	1207 9
brie	308 50	1196 16	234 57	1627 22	275 37	2467 25
dri	380 26	1811 28	226 54	1830 28	277 34	2216 18
gris	302 30	2106 18	232 44	2040 26	275 33	2109 13
brou	246 57	724 34	141 53	522 18	192 51	574 23
droue	290 37	910 26	180 35	566 30	261 47	633 35
groux	285 38	642 30	206 43	632 23	203 39	612- 32

Tableau 9: moyennes (en Herz) et coefficients de variation des valeurs de /r/."a" représente les valeurs de F1 (basses fréquences) et "b" les valeurs de F2; "voc1", "bat", et "voc 2" signifient respectivement que les mesures ont été prises au centre de la première partie vocalique, au centre du battement et au centre de la deuxième partie vocalique.

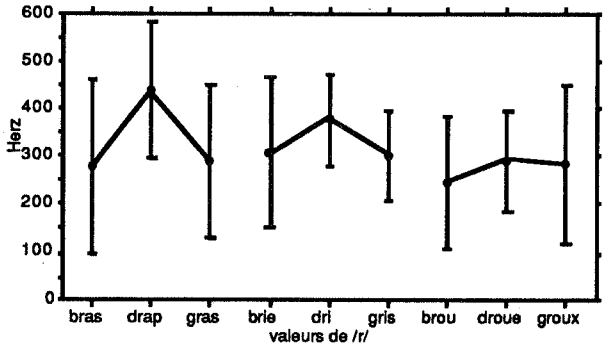


Figure 8: moyennes et écarts types des valeurs du F1 de /r/ mesurées dans la première partie vocalique de la consonne (voc1). En abcisse les fréquences en Herz, en ordonnée les contextes de réalisation. Nous avons relié par un trait continu les contextes vocaliques identiques.

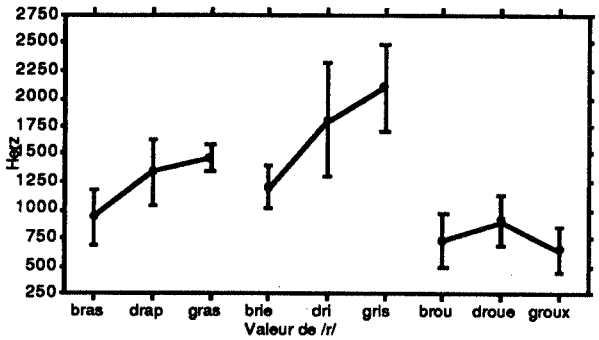


Figure 9: moyennes et écarts types des valeurs du F2 de /r/ mesurées dans la première partie vocalique de la consonne (voc1). En abcisse les fréquences en Herz, en ordonnée les contextes de réalisation. Nous avons relié par un trait continu les contextes vocaliques identiques.

Le /r/ subit l'influence de C1 dans sa première partie vocalique. Il est intéressant de constater que F1 et F2 subissent une influence différente (figures 8 et 9). Alors que la voyelle adjacente fait varier F2, elle semble jouer un rôle mineur, sinon inexistant, dans dans les variations de F1. Nous constatons que la vélaire /g/ entraîne une variation particulière sur /r/ lorsque la voyelle adjacente est /u/. Ce phénomène très intéressant (déjà observé sur la figure 6), peut être interprété de la façon suivante: dans une suite CV, le changement de voyelle provoque la réalisation de différents allophones de le consonne /g/; ainsi, dans une suite CCV, l'influence de la voyelle sur C1 semble se répercuter sur C2. Cette observation sur C2 est d'autant plus intéressante qu'elle nous informe sur la nature de C1, l'occlusive, qui est difficilement identifiable par la nature de son explosion.

6-CONCLUSIONS

Bien qu'il semble difficile d'élaborer des considérations générales concernant la dimension l'acoustique des consonnes, nous pensons qu'il est possible de dégager certaines tendances de l'étude que nous avons effectuée.

En premier lieu, nous constatons que la variation porte le plus souvent sur la partie qui se trouve au contact de la consonne porteuse de variabilité et non sur l'ensemble du segment consonantique étudié. On constate toutefois que, dans certains cas (/l/), la variation s'étend jusqu'au centre du segment mais de façon très atténuée.

Le F2 subit plus souvent la variation que le F1. Il est cependant intéressant de constater que lorsque le F1 est influencé (figure 8), il est possible que cette variation soit différente de celle du F2 (figure 9). On peut penser ici à un rôle éventuel du locus des occlusives en fonction des voyelles de la syllabe.

Il est difficile d'appréhender le sens de la variation. On sait que la voyelle infuence les unités phonétiques qui la précèdent. On sait également que C1 influence C2 et inversement. Il semble qu'au sein du GC l'influence de C1 vers C2 (figure 2) soit plus importante que l'influence de C2 vers C1 (tableaux 1 et 2). On doit toutefois nuancer ce propos car, dans ce cas, l'observation de la variation ne porte pas sur la même unité phonétique; il est envisageable que chaque unité phonétique ne subisse pas la variation de la même façon.

On constate que la voyelle adjacente influence l'ensemble du GC (Meunier, 1992). Il est cependant intéressant de noter que le rôle de la voyelle ne masque pas celui de la consonne porteuse de variabilité (figures 2, 3, 6 et 9).

Enfin, nous évoquerons la complexité du phénomène de la variation en rappelant le comportement du /r/ dans les groupes Hé3. On constate dans ce cas que l'influence de la voyelle sur les occlusives /k/-/g/ se répercute sur la consonne vocalique intermédiaire: autrement dit, la consonne vocalique conserve la valeur des transitions de l'occlusive précédente (/k/ /g/) en fonction de la voyelle suivante.

REFERENCES

AUBERGE, V., et al., 1988,"Lexique et groupes consonantiques", *Actes des 17èmes Journées d'Etude sur la Parole*, Nancy, 55-60.

AUTESSERRE, D., ROSSI, M., 1987, "La segmentation et l'étiquetage des groupes consonantiques de la BDSON", *Actes des 16èmes Journées d'Etude sur la Parole*, Hammamet, 196-199.

CARRE, R. et al., 1987, "La base de données des sons du français (BDSONS). Perspectives et développement", *Actes des 16èmes Journées d'Etude sur la Parole*, Hammamet, 335-337.

HJELMSLEV, L., 1963, *Le Langage*, coll. Folio Essai, Gallimard.

MARCHAL, A., 1985, "La coproduction dans les groupes d'occlusives", *Actes des 14èmes Journées d'Etude sur la parole*, 19-22.

MEUNIER, C.,1990, "Groupes consonantiques: premier inventaire des réalisations acoustiques des phases de transition", *Actes des 18èmes Journées d'Etude sur la Parole*, Montréal, p. 69-73

MEUNIER, C.,1992, "Variabilité acoustique des groupes de consonnes : le rôle de la voyelle adjacente"*Actes du 12ème Congrès Français d'Acoustique*, Arcachon, Avril 1992. (A paraître).

PULGRAM, E., 1965, "Consonant cluster, consonant sequence, and the syllabe.", *Phonetica*, 13, 1-2, 76-81.

ROCHETTE, C., 1973, *Les Groupes de consonnes du français: étude de l'enchaînement articulatoire à l'aide de la radiocinématographie et de l'oscillographie*, Klincksieck, Québec-Paris (?).

ROSSI, M., 1968, "Au sujet des groupes consonantiques du français", *Revue d'Acoustique*, 3, 4, 306-311.

19es J.E.P. - Bruxelles, 19 au 22 mai 1992.

Extraction des traits distinctifs par un réseau neuronal

Shigeyoshi KITAZAWA*, Yukihiro NISHINUMA** & Takahiko SHINMURA*

Université de Shizuoka, 3-5-1 Johoku, Hamamatsu, Japon *
Université de Provence, CNRS U.R.A.-261, 13621 Aix-en-Provence, France **

Résumé.
Ce travail concerne le réseau neuronal que nous avons expérimenté pour extraire les traits distinctifs des voyelles. Le corpus utilisé comporte 5 voyelles: /i, e, a, o, u/ et 30 syllabes du type CV avec 6 consonnes plosives: /p, t, k, b, d, g/ réalisées par 82 sujets japonais. Nous avons d'abord effectué une analyse LPC sur le segment de la voyelle découpée en 15 trames de largeur égale; ensuite sur le spectre de puissance obtenu par FFT nous avons appliqué la courbe isosonique (dB) de façon à simuler le spectre auditif réparti en 22 bandes critiques. Le réseau neuronal utilisé comprend 2 couches cachées. Son apprentissage et le test ont été effectués sur les 43050 données spectrales. Les résultats démontrent que le réseau a extrait les caractéristiques acoustiques pour détecter les oppositions des traits: compact/diffus, grave/aigu et bémolisé/ non-bémolisé. La performance globale est comparable à celle de la fonction discriminante linéaire obtenue sur les mêmes données.

1. Introduction.

La reconnaissance automatique de la parole consiste en une exploitation systématique des caractéristiques physiques observables dans le signal parlé, auxquelles on donne différents termes: indices acoustiques, traits distinctifs, etc. L'origine de la notion de traits distinctifs remonte assez loin dans l'histoire linguistique, mais c'est sans aucun doute les travaux de Jakobson, Fant et Halle qui ont marqué l'époque; ils ont défini initialement leurs traits au niveau articulatoire, acoustique et perceptif [1]. Ces auteurs imaginaient une pondération du spectre par une fonction auditive pour établir une correspondance entre les niveaux acoustique et perceptif des traits. Malheureusement l'état de l'art d'alors ne permettait pas d'y parvenir. C'est probablement pour cette raison que dans la phonologie générative, l'aspect physiologique des traits a prédominé pendant longtemps. Après de multiples mouvements évolutifs bien connus, les indices et les traits sont actuellement mieux hiérarchisés, de plus l'aspect acoustique reprend une place valorisée [2]. D'autre part, indépendamment de cette école phonologiste, Ladefoged et Halle ont aussi défini les traits du point de vue articulatoire et acoustique [3]. Par ailleurs, en vue de la reconnaissance automatique de la parole, l'importance de la réalité acoustique et perceptuelle a été soulignée dans le cadre de la théorie des traits distinctifs [4]. En somme il existe un mouvement de retour à la source jakobsonienne.

La reconnaissance de la parole est assimilable aux problèmes de classification des objets en général; en ce qui concerne la méthodologie de cette dernière, la statistique multivariable se montre sûre depuis de longues années et la technique neuronale a commencé à apporter une contribution significative [5]. De même dans notre communauté de la parole, l'analyse discriminante est un outil de travail indispensable et nous en avons profité utilement [6; 7]. Les réseaux neuronaux sont aussi de plus en plus expérimentés par des chercheurs de différentes formations [8; 9]. Toutefois, dans la plupart de ces travaux, les réseaux neuronaux sont utilisés comme classificateur final des phonèmes. En revanche nous nous intéressons plus particulièrement au système neuromimétique en tant que mécanisme

extracteur des traits (ceux-ci seront ensuite utilisés pour une classification définitive par un système de connaissances). Si le fonctionnement d'un réseau neuronal se trouve satisfaisant, l'examen de la fonction de transfert entre les couches du réseau peut nous révéler des faits enrichissants sur la parole au même titre que l'examen des renseignements détaillés des analyses statistiques multivariables. Ces deux techniques semblent précieuses et sont complémentaires de l'expert humain qualifié.

Dans les pages qui suivent nous allons décrire notre expérience, dans laquelle les données acoustiques ont été d'abord pondérées suivant la perception d'intensité avant d'alimenter le réseau neuronal. On examinera les résultats obtenus sur le réseau neuronal, dans son comportement normal et ses erreurs, qui seront comparés avec ceux de l'analyse discriminante.

1. 1. Corpus.

Si le nombre de voyelles est un indicateur de complexité pour une langue, le français ou l'anglais se placent à l'opposé du japonais qui est un système plutôt simplifié avec ses 5 voyelles. C'est ainsi qu'avant d'examiner les données du français, nous avons testé notre démarche sur le japonais. Le corpus utilisé comporte donc 5 voyelles: /i, e, a, o, u/ et 30 syllabes du type CV (consonne + voyelle) avec 6 consonnes plosives: /p, t, k, b, d, g/. La spécification des traits de ces voyelles est indiquée dans le Tableau 1. Ce matériel a été lu et enregistré dans une chambre anéchoïque par 82 sujets masculins japonais.

Oppositions	/o/	/a/	/e/	/u/	/i/
compact/diffus	+	+	+	-	-
grave/aigu	+	+	-	+	-
bémolisé/ non-bémolisé	+	-		+	

Tableau 1
Traits définissant les 5 voyelles japonaises.

1.2. Analyse acoustique.

Le signal a été échantillonné à 16 kHz. On a considéré comme voyelle la zone de CV dont l'énergie RMS est supérieure à 18% de l'énergie maximale de la portion CV. Cette voyelle a été découpée en 15 trames; pratiquement, une fenêtre de 256 points (= 16 ms) correspondant à une trame a été déplacée d'un pas régulier du début jusqu'à la fin de la voyelle. Sur chaque trame nous avons effectué une analyse LPC (type de fenêtre utilisée: Hamming) avec 23 coefficients de façon à extraire les caractéristiques de la fonction de transfert. Le spectre de puissance lissé (dB) a été obtenu par FFT. Ce spectre a été ensuite découpé en 22 bandes critiques en vue d'une pondération par rapport à la courbe d'isosonie (niveau sonore, 60 dB; cf Figure 1) [10; 11]. Nous avons ainsi simulé le spectre auditif au sens fletchérien du terme et pour répondre à la question posée par Jakobson et al. (voir ci-dessus). Le nombre total des données (trames) s'élève à 43050 (15 trames * 35 syllabes * 82 sujets).

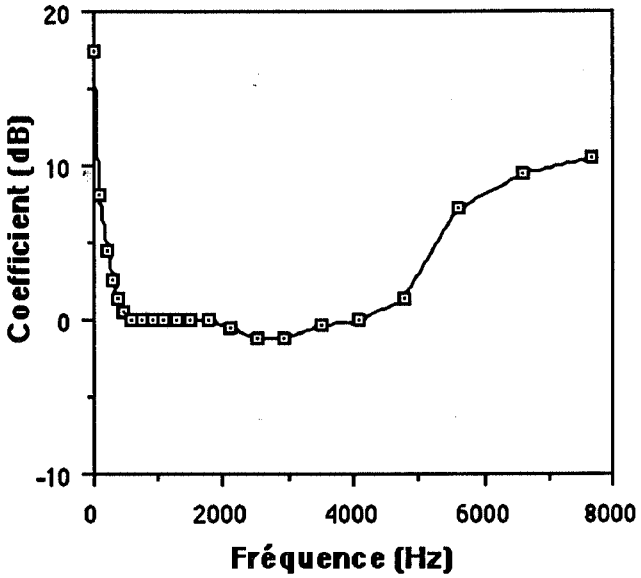


Figure 1
Courbe d'isosonie utilisée.

1. 3. Réseau neuronal.

Le réseau neuronal que nous avons utilisé est du type perceptrons à couches multiples comme le montre la Figure 2. Trois apprentissages avec différentes configurations de couches nous ont permis de choisir cette meilleure structure neuronale. En utilisant 4200 trames prises au hasard parmi 43050, nous avons effectué l'apprentissage de l'opposition compact/ diffus; constatant une saturation aux alentours de 800 répétitions, nous avons arrêté l'apprentissage au bout de 1000 (cf. Tableau 2). La couche d'entrée comprend 22 cellules pour les données

spectrales décrites plus haut, la couche cachée 1 compte 35 cellules (hypothèse liée au nombre de contextes syllabiques), la couche cachée 2 comprend 5 cellules (lien hypothétique avec les voyelles) et enfin la couche de sortie 2 cellules. La cellule notée B dans la figure désigne "bias term" dont la valeur est constante de 1.

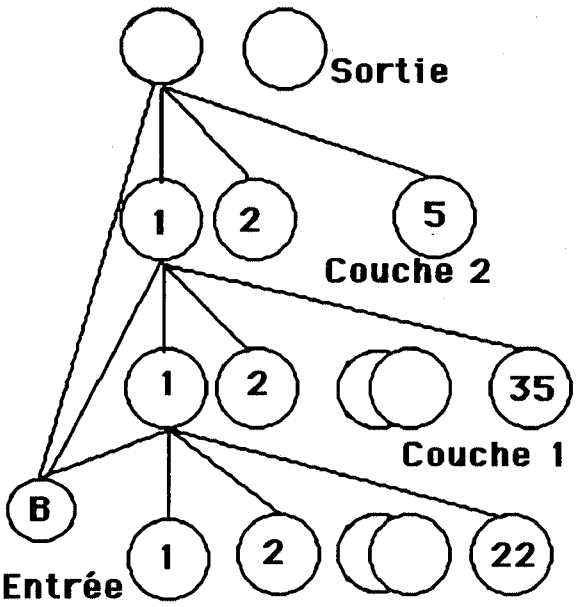


Figure 2
Structure du réseau neuronal utilisé.

Architecture	I	II	III
Couche d'entrée	22	22	22
Couche cachée 1	35	35	256
Couche cachée 2	0	5	0
Couche de sortie	2	2	2
Connections totales	877	997	6402
Apprentissage (heure)	26	33	253
Taux de réussite (%)	96,8	96,9	96,8

Tableau 2
Comparaison des structures du réseau neuronal

Pour faire fonctionner ce système neuronal, nous avons eu recours à un algorithme dérivé de la méthode DCP et réalisé dans DCP2, logiciel développé par ATR.

2. Résultats et Discussion.

La discussion qui suit porte sur les résultats obtenus de ce réseau neuronal qui a été de nouveau entraîné et testé avec la totalité des données. Les résultats récapitulatifs de classification en terme de trait et d'opposition sont résumés au Tableau 3.

	trait	opposition
Compact	96,3%	
Diffus	95,8%	96,1%
Grave	96,9%	
Aigu	94,9%	95,9%
Bémolisé	99,0%	
Non-bémolisé	94,6%	97,5%

Tableau 3
Détection des traits par le réseau neuronal.

D'une manière générale, le comportement du système est correct avec un taux moyen de détection des traits atteignant 96,3%. Le trait "bémolisé" est celui qui a été le mieux détecté, en revanche sa contrepartie "non-bémolisé" est celle qui a le moins bien fonctionné.

Le Tableau 4 nous fait remarquer qu'un trait peut produire des résultats négatifs plus ou moins importants suivant les voyelles. Si le trait "bémolisé" pour /u/ a connu les erreurs les moins nombreuses (66), "aigu" pour /i/ (646) ou "grave" pour /u/ (598) en ont cumulé dix fois plus. De même les traits compact et grave pour /a/ comportent moins d'erreurs (90 et 96 respectivement) que le trait non-bémolisé (462).

	i	e	a	o	u
Compact		495	90	375	
Diffus	324				393
Grave			96	113	598
Aigu	646	310			
Bémolisé				110	66
Non-bémolisé			462		

Tableau 4
Erreurs de détection par trait et par voyelle.

En observant la distribution des poids calculés pour la fonction de transfert de la couche d'entrée à la couche cachée 1, nous avons remarqué une valeur forte du poids pour les zones fréquencielles correspondant aux deux premiers formants; par exemple, entre 500 Hz et 1500 Hz quend il s'agit du trait compact; entre 800 Hz et 1800 Hz pour le trait grave ainsi

qu'entre 300 Hz et 600 Hz pour le trait bémolisé. Nous pensons donc que le réseau neuronal tente d'extraire les caractéristiques physiques définissant chaque trait.

En ce qui concerne la localisation des erreurs de détection dans le temps, on note d'abord que la moitié des mauvais scores se concentre sur les trois premières trames (trames 0, 1 et 2), plus particulièrement la trame 0 détient plus de 30% d'erreurs comme le montre la Figure 3. Ici, vraisemblablement, l'effet de co-articulation a dû perturber le comportement de notre réseau neuronal. La seconde moitié des erreurs s'étalant sur le reste du segment, c'est-à-dire de la trame 3 à la trame 14, peut indiquer d'autres oppositions des traits dont on ignore l'identité.

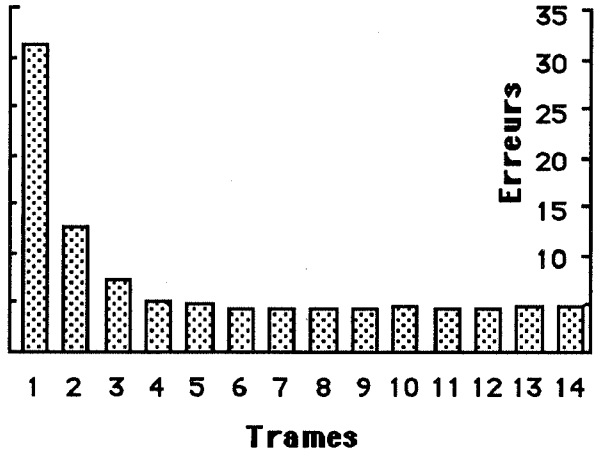


Figure 3
Erreurs observées en fonction des trames.

	Perceptrons		Discriminante	
	Trait	Moy.	Trait	Moy.
Compact	96,3		94,9	
Diffus	95,8	96,1	96,5	95,6
Grave	96,9		96,0	
Aigu	94,9	95,9	96,0	95,6
Bémolisé	99,0		98,1	
Non-bémolisé	94,6	97,5	94,5	96,9

Tableau 5
Résultats du réseau neuronal et de la fonction discriminante.

Pour savoir dans quelle mesure la solution neuronale est efficace, nous avons effectué une analyse discriminante sur les mêmes données. Le Tableau 5 met en évidence le fait que les deux modèles ont fourni des résultats quasiment identiques. En effet, même si le réseau neuronal fonctionne très légèrement mieux, la différence entre les deux ne dépasse pas 0,3%.

3. Conclusion.

Afin d'améliorer notre système de reconnaissance automatique de la parole, nous avons examiné la possibilité d'utiliser un réseau neuronal en tant qu'extracteur des traits distinctifs des voyelles. Les résultats obtenus laissent penser qu'effectivement les perceptrons cherchent à repérer les indices physiques des traits mentionnés par Jakobson et al.

Compte tenu de la nature des erreurs, l'effet de co-articulation sur le comportement de notre réseau neuronal ne semble pas d'une quantité négligeable pour un système de reconnaissance exigeant. Toutefois le système neuronal artificiel ayant montré une performance comparable sinon supérieure à une fonction discriminante, il peut être d'une grande utilité pour enrichir nos connaissances sur la parole.

Références

[1] JAKOBSON, R., FANT, C. G. M. & HALLE, M., *Preliminaries to Speech Analysis: The Distinctive Features and Their Correlates*, MIT Press (1951).

[2] SAGEY, E., *The Representation of Features and Relations in Non-linear Phonology*, PhD diss. (1986).

[3] LADEFOGED, P. & HALLE, M., Some major features of the International Phonetic Alphabet, *Language* **64** (1988) 577-582.

[4] DANTUJI, M., A preliminary study on a new acoustic feature model and feature hierarchies, *Proc. First ROC- Japan Seminar on New Speech Recognition Method* (1991) 151-165.

[5] LIPPMANN, R. P., An Introduction to Computing with Neural Networks, *IEEE ASSP Magazine*, (1987).

[6] KITAZAWA, S. & TUBACH, J. P., Discriminant analysis and perceptual test of french stops and nasals", *Proc. 9th Int. Conf. Pat. Recog.* (1988) 1077-1079.

[7] NISHINUMA, Y. DUEZ, D; & PABOUDJIAN, C., "Automatic classification of consonant clusters in French", *Speech Communication*, **10** (1991) 395-403
[8] LEUNG, H. C. & ZUE, V. W., Phonetic Classification using multi-layer perceptrons, *Proc ICASP-90* (1990) 525-528.
[9] MENG, H. & ZUE, V. W., A comparative study of acoustic representations of speech for vowel classification using multi-layer perceptrons, *Proc ICSLP -90* (1990) 1053-1056.
[10] FLETCHER, H., *Speech and Hearing in Communication*, D. Van Nostrand, (New York, 1953)
[11] ISO-Rec. 226, Equal Loudness Contours for Pure Tones (Free Field), Threshold of Hearing (Binaural Free Field), Age Corrections.

19es J.E.P. - Bruxelles, 19 au 22 mai 1992.

LE BE - BEGAYAGE ET EUH...., L'HESITATION EN FRANCAIS SPONTANE

BRIGITTE ZELLNER-BECHEL

LABORATOIRE DE PHONETIQUE, U.F.RECHERCHES LINGUISTIQUES, UNIVERSITE PARIS 7

Résumé

To clarify what is normal speech vs abnormal speech, the purpose, in this preliminary study, is to define hesitation and stuttering. As these two verbal behaviours may seem to be similar because of the disfluencies (the disorders of fluency), in particuliar when the stutterer is a mild stutterer, the question is:

- Can hesitation and stuttering be differentiated by analysing the verbal productions ?

The "cascade effect", the absence of "the effect of fonction words", the syllabic dislocations seem to be good predictors for stuttering. An attempt to formalize a procedure to differentiate these two verbal behaviours is then proposed.

INTRODUCTION

L'étude de la parole spontanée présuppose que l'objet même de la recherche soit clairement défini. C'est ainsi que les domaines de la parole "normale" et de la parole "pathologique" doivent être, autant que faire se peut, délimités. Or si la nosologie des cas pathologiques graves est aujourd'hui assez bien établie, (aphasies, apraxies), nous ne savons toujours pas comment s'effectue le passage du "normal" à l'"anormal". C'est dans cette perspective que se situe cette étude pour tenter de définir, au plan de la production de la parole, quelles sont les manifestations de l'hésitation et du bégayage (ou acte de bégayer).

L'hésitation peut se définir comme un comportement d'indécision, et en particulier comme un comportement d'indécision verbale.

"J'hésite par exemple, / v / **euh** lorsque je dois
Prolongation **Répétition**
choisir **un**--------- **un** mot, une
expression".

Le bégaiement peut se définir comme un trouble pathologique du comportement verbal. On diagnostique le bégaiement lorsqu'en particulier, la fluidité de la parole est atteinte. Le bégaiement se manifeste donc par des bégayages.

2 Répétitions du mot
"Le bégayage survient de **de de** manière
2 Répétitions du phonème
im **im** **im**prévisible et se caractérise par
une incapacité involontaire à dire".

QUESTION

Bien qu'hésitation et bégayage se ressemblent dans leurs manifestations (ie: les accidents de parole du type répétitions de mots, prolongations de syllabes, faux départs, etc. Cf Wingate 1987), est-il possible de différencier le comportement d'un sujet hésitant de celui d'un sujet bègue, au plan de la production de la parole ?

REMARQUES

1/ C'est l'analyse des productions verbales qui doit permettre de définir comment se manifestent hésitation et bégayage: toute définition préalable à cette analyse a donc été rejetée.

2/ Pour ce faire, tous les événements du type bredouillements, prolongations, faux départs, bégayages,

etc, ont été considérés comme des accidents de parole.

METHODOLOGIE

22 locuteurs adultes, de langue maternelle française, dont 11 sujets bègues et 11 sujets non bègues, ont été soumis à un test en images (situation de langage spontané contrainte). En moyenne, les productions orales des sujets bègues comprenaient 154 mots, celle des sujets non bègues 138 mots. Les corpus ont été transcrits selon l'adaptation au français du "Systematic Disfluency Analysis" de Campbell et Hill, (Cf Zellner-Béchel et Pfauwadel; 1991). Ce système de transcription, qui permet de fonder une procédure d'analyse du bégaiement, présente plusieurs avantages.

Tout d'abord, il permet d'observer le comportement verbal dans son ensemble -avant, pendant et après les accidents- donnant ainsi une **dimension temporelle** du bégaiement que l'on ne saurait obtenir par le simple comptage des accidents de parole. Ensuite, il prend en compte la notion de "continuum d'accidents de parole", mise en évidence par Gregory et Hill (1980): de la simple hésitation au bégayage franc, il n'y a pas passage d'une qualité d'accidents à une autre; et c'est bien là toute notre problématique. Enfin c'est un système de transcription qui semble fiable. A quelques variations près, on obtient des transcriptions similaires:

-avec un autre auditeur respectant la même procédure;

-après analyse instrumentale à l'aide des logiciels d'analyse du signal de parole *Sound Edit™* et *Signalyze™*.

A partir de ces transcriptions, le nombre de mots subissant un/des accidents et le nombre d'accidents ont été calculés en fonction du nombre de mots produits par chaque locuteur selon leur lieu d'occurrence, leur catégorie lexicale et leur longueur syllabique. L'analyse syntaxique et morphologique des productions orales ne pouvant être parfaitement décrite avec les outils de la langue écrite, il s'en suit que les erreurs de catégorisation étaient inévitables. Pour les minimiser et les homogénéiser, une seconde analyse des corpus a été effectuée, permettant de vérifier la concordance des deux études.

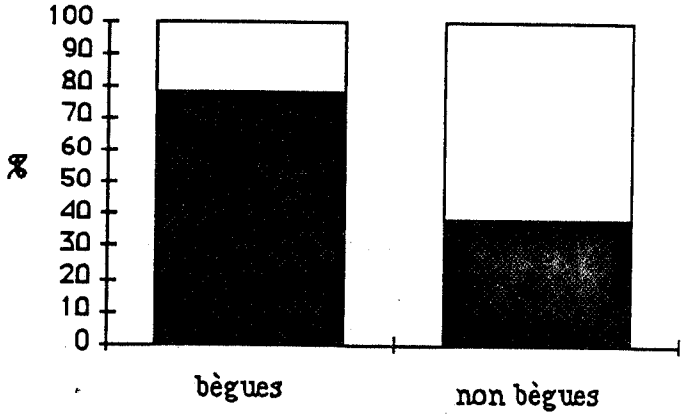
RESULTATS

Il faut remarquer tout d'abord qu'il n'est pas apparu de différences statistiques significatives entre les deux populations du point de vue de la quantité des mots produits et de leur répartition en mots fonction (déterminants, pronoms, conjonctions, prépositions, interjections) et mots lexicaux (noms, verbes, adjectifs, adverbes).

La différenciation des deux comportements peut s'établir aux trois niveaux: syntaxique, lexical, syllabique.

1/ Au passage d'une frontière syntaxique (Cf Boulakia, 1983; Klouda et Cooper, 1987), ce sont les "cascades d'accidents de parole" succédant à une pause silencieuse qui révèlent une tendance au comportement bègue. Alors qu'une tendance au comportement hésitant se manifeste plutôt par la production d'accidents isolés, sans interruption du flux sonore.

Tableau 1: Proportion d'accidents en cascade (en noir) et d'accidents isolés (en blanc) dans l'ensemble des accidents



La différence entre les deux populations est particulièrement significative:

chi2 = 65,962; p≤ 0,0001; df=1.

2/ Au niveau lexical (Cf Vaane, 1979), du fait de la grande variabilité des comportements, il était intéressant d'étudier **l'évolution** des variations entre:

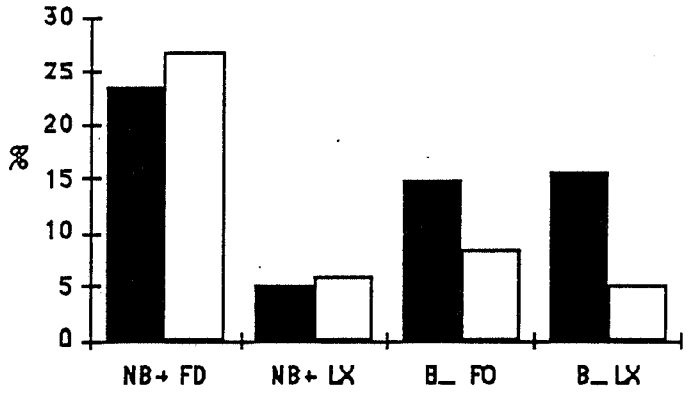
- d'une part les sujets bègues "les moins bègues" (B-) et les sujets bègues "les plus bègues" (B+),
- d'autre part les sujets "les moins hésitants" (NB-) et les sujets "les plus hésitants" (NB+),
- enfin, les sujets "les plus hésitants" (NB+) et les sujets bègues "les moins bègues"(B-)

Il apparait que, quelle que soit la gravité du bégaiement, lors d'une phase de bégayage, et compte tenu du phénomène des cascades d'accidents, TOUS les mots sont susceptibles de subir un accident de parole au moins. (L'écart entre la fréquence des accidents sur mots fonction et celle sur mots lexicaux n'est pas significatif: chi 2= 2,187; p= 0,14; df= 1.)

En revanche, lors d'une hésitation, quel que soit le taux de disfluences, les mots fonction monosyllabiques -qui permettent de structurer le discours-, ont tendance à subir plus d'accidents relativement aux autres mots. (L'écart entre la fréquence des accidents sur mots fonction et celle sur mots lexicaux est significatif:

chi2= 137,061; p≤ 0,0001; df= 1.)

Tableau 2: Pourcentage des accidents en fonction des mots de même longueur et de même catégorie, de NB+ à B-



Légende: - en noir: mots 1 syllabe
- en blanc: mots > 1 syllabe
FO: mots fonction
LX: mots lexicaux

Les différences majeures entre NB+ et B- portent donc sur :
accidents sur mots fonction,
et accidents sur mots lexicaux .

3/ Au niveau syllabique la structure interne des mots est également atteinte lors d'une phase de bégayage. Il existe à ce niveau deux types d'accidents:

a) LA DISLOCATION DE SYLLABE
Dans un mot, une syllabe est: a) soit progressivement produite en plusieurs émissions; b) soit produite avec un rapport temporel consonne/voyelle anormal.

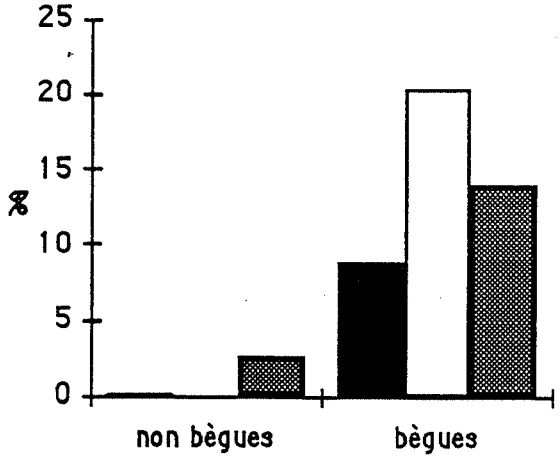
Exemple: "g_ g ga ar ç çon"

b) L'ACCIDENT AVEC SYLLABATION
Dans un mot, l'accident survient exactement à la frontière syllabique.

Exemple : " main **euh** tenant"

Ces deux types d'accidents représentent 40 % des disfluences produites par les bègues. Tandis que lors d'une hésitation, la structure syllabique est respectée.

Tableau 3: Proportion des accidents syllabiques dans l'ensemble des accidents sur mots de même longueur



Légende: en noir: dislocations sur mots 1 syllabe
en blanc: dislocations sur mots >1 syllabe
en gris: syllabations sur mots >1 syllabe

Finalement, l'analyse à ces trois niveaux permet d'établir que:
-il y a **bégayage** lorsque le locuteur **perd momentanément le contrôle** de sa production verbale aux niveaux syntaxique et/ou lexical et/ou syllabique.
-il y a **hésitation** lorsque le locuteur **conserve le contrôle** de sa production verbale aux niveaux syntaxique et/ou lexical et/ou syllabique.

FORMALISATION

La PROCEDURE DE DIFFERENCIATION des

deux comportements peut alors se formaliser comme suit.

I / NIVEAU SYNTAXIQUE

1/ Calculer le taux x1 d'accidents en cascades.

2/ Calculer le taux x2 de cascades précédées d'une pause silencieuse.

3/ Pour ces deux taux, appliquer la formule suivante:

SI $x > 50\ \%$ ALORS $A = 1$
SI $x = 0\ \%$ ALORS $A = 0$
SINON $A = 0{,}5$

4/ Calculer

$$A = A1 + A2$$

II / NIVEAU SYLLABIQUE

1/ Calculer le taux y de dislocations.

2/ Appliquer la formule suivante:

SI $y > 10\ \%$ ALORS $B = 1$
SI $y = 0\ \%$ ALORS $B = 0$
SINON $B = 0{,}5$

III / ANALYSE DES DEUX NIVEAUX

1/ Calculer A+B=

2/ Analyse:

SI $A + B > 2$

Interprétation: il y a une perte IMPORTANTE du contrôle de la production verbale aux niveaux syntaxique et/ou syllabique.

ALORS **il s'agit d'un comportement typiquement bègue.**

SINON

SI $A \geq 1$ et $B \geq 1$

Interprétation: le locuteur PERD le contrôle de sa production verbale au niveau syntaxique ET au niveau syllabique.

ALORS **il s'agit d'un comportement typiquement bègue.**

SINON

SI $A \leq 0{,}5$ et $B \leq 0{,}5$

Interprétation; le locuteur CONSERVE le contrôle de sa production verbale au niveau syntaxique ET au niveau lexical.

ALORS **il s'agit d'un comportement typiquement hésitant.**

SINON continuer l'analyse.

IV/ NIVEAU LEXICAL

1/ Calculer l'écart-type σ entre la moyenne d'accidents sur mots fonction monosyllabiques et la moyenne d'accidents sur mots lexicaux monosyllabiques.

2/ SI $\sigma \leq 1$

Interprétation: TOUS les mots sont susceptibles de subir un accident

ALORS **il s'agit d'un comportement typiquement bègue.**

SINON

Interprétation: tendance à produire PLUS D'ACCIDENTS sur mots fonction monosyllabiques.

ALORS **il s'agit d'un comportement typiquement hésitant.**

CONCLUSION

Au plan de la production du langage, la formalisation du comportement bègue et du comportement hésitant permet de mettre en évidence que:

- le bégayage est l'expression d'un comportement PASSIF; le locuteur bégayant perd momentanément le contrôle de sa production verbale aux niveaux syntaxique, lexical et syllabique.
- L'hésitation est l'expression d'un comportement ACTIF; le locuteur hésitant conservant le contrôle de sa production verbale aux trois niveaux.

REFERENCES BIBLIOGRAPHIQUES

BOULAKIA,G. (1983), "Phonosyntaxe du français", *T.A. Information, 24-2, p.24-63.*

CAMPBELL, J.H. & HILL, D.G. (1987), "Systematic Disfluency Analysis", *Acts of Department of Communication Sciences and Disorders Speech and Language Clinic*, Northwestern University.

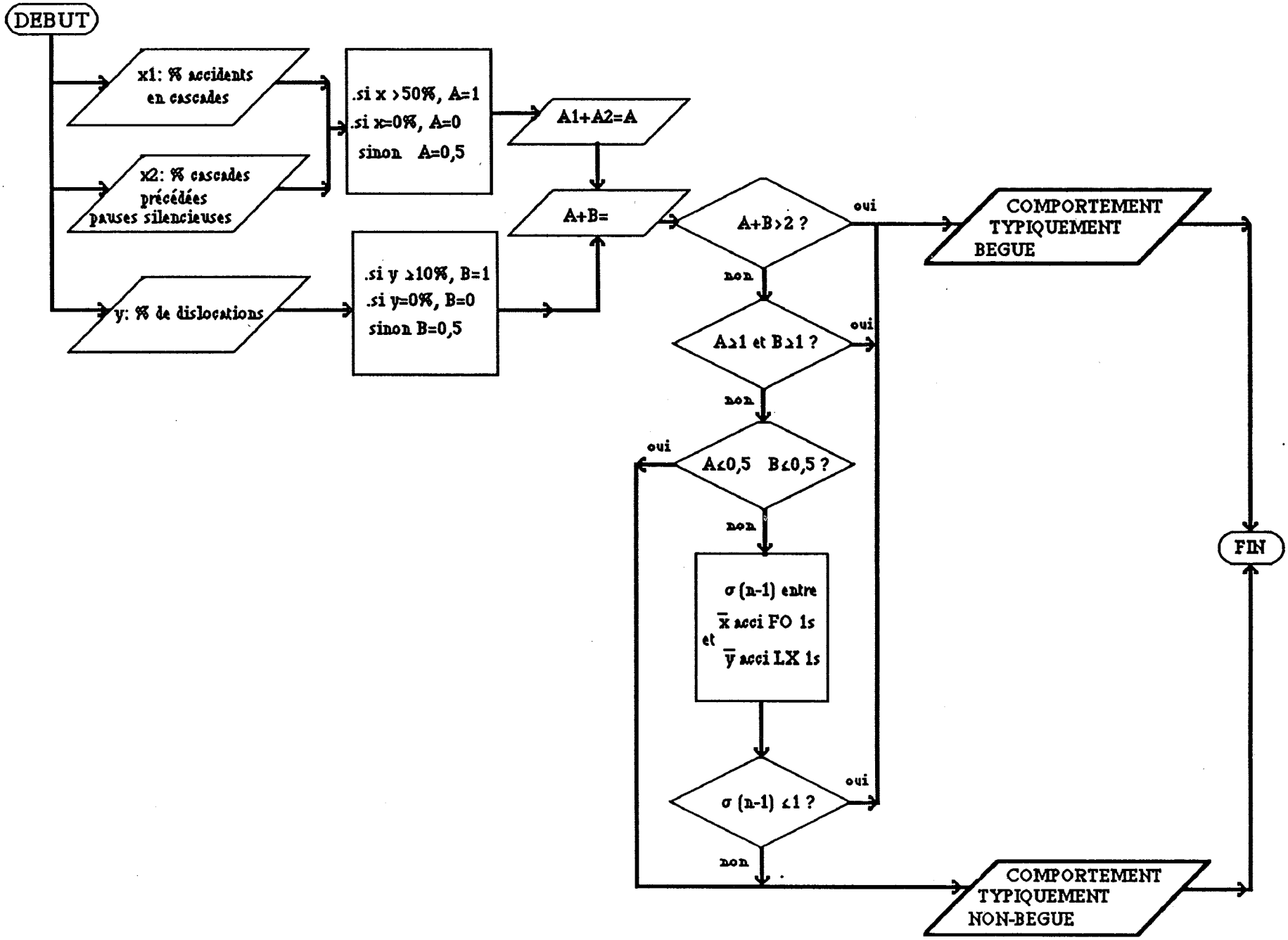
KLOUDA, G. & COOPER, W. (1987), "Syntactic Clause Boundaries, Speech Timing and Stuttering Frequency in Adult Stutterers", *Language and speech, 30-3.*

VAANE, E. (1979), "Disfluencies in Speech and some Linguistic Factors: a Comparison between Stuttering and non Stuttering Persons", *P.R.I.P.U, 4 (2).*

WINGATE, M. (1987), "Fluency and Disfluency; Illusion and Identification" *J Fluency. Dis., 12-1.*

ZELLNER-BECHEL, B & PFAUWADEL, M.C. (1991), "Adaptation du *Systematic Disfluency Analysis* au français", conférence internationale sur le bégaiement, Saulx Les Chartreux.

Organigramme: Procédure de différenciation des deux comportements



19es J.E.P. - Bruxelles, 19 au 22 mai 1992.

ETUDE FORMANTIQUE DES VOYELLES DE L'ARABE STANDARD (LOCUTEURS : UN MAROCAIN, UN ALGÉRIEN ET UN TUNISIEN)

I. ZNAGUI

INSTITUT DE PHONÉTIQUE (URA1027)
UNIVERSITÉ DE LA SORBONNE NOUVELLE 19 RUE DES BERNARDINS 75005 PARIS

Résumé

ABSRACT

F1 and F2 of arabic vowels preceded by consonants of different places of articulation were studied. The first formant was found to be higher in vowels preceded by emphatic, uvular and pharyngeal consonants than in vowels preceded by labial and dental consonants. Statistically, two classes of vowels were distinguishable: 1- vowels with a high F1. 2- vowels with a low F1. A binary classification of arabic vowels for speech synthesis can be best based upon the first formant, being higher or not.

Introduction

Les travaux antérieurs sur le système consonantique et vocalique de l'arabe se sont surtout intéressés au phénomène d'emphase ou de pharyngalisation [1,2].

Aussi, avons-nous pensé à analyser les formants des voyelles de l'arabe standard ou "moderne" précédées de toutes les consonnes. Dans une première étape, cette étude a été réalisée avec trois locuteurs maghrébins. Dans la variabilité formantique de F1 et F2 des voyelles de ces trois locuteurs, nous avons constaté des points communs.

A notre avis, l'élévation de F1 ne serait pas spécifique des voyelles précédées d'une consonne emphatique, réalisée avec un recul de la racine de la langue vers la paroi pharyngale, mais elle constituerait un indice acoustique qu'on retrouverait dans les consonnes uvulaires réalisées avec la luette et les consonnes pharyngales (2,3) (voir tableau des consonnes : figure 1).

Figure 1

Le système consonantique de l'arabe standard

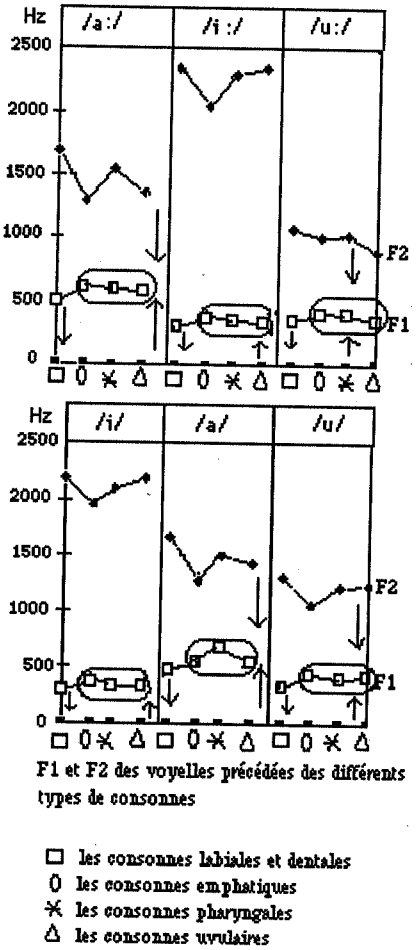
Les points d'articulation

		Bilabiale	labiodentale	interdentale	dentale	palatale	post-palatale	vélo-uvulaire	pharyngale	glottale
Nasales		m			n					
orales		b	f	θ	t					
		w		ð	d					
					s					
					z					
					ʃ					
					l					
					r					
						j				
							k			
								q		
								x		
								ɣ		
									ʕ	
									ħ	
										ʔ
										h
les consonnes emphatiques					tˤ					
					dˤ					
					sˤ					
					zˤ					
					ɾˤ					
					lˤ					

Méthodologie

Nous avons constitué un corpus de voyelles précédées des 28 consonnes arabes dans des mots réels polysyllabiques intégrés dans des phrases cadres: "qa:la *sabi:lun* marratan" ("il a dit [chemin]une fois"). Trois locuteurs : un marocain, un algérien et un tunisien ont répété trois fois ces phrases . Les mesures formantiques ont été faites au centre des voyelles à l'aide du programme Mac Speech Lab , le signal ayant été échantillonné à une fréquence de 10Khz avec une dynamique de 12 bits.

Observations

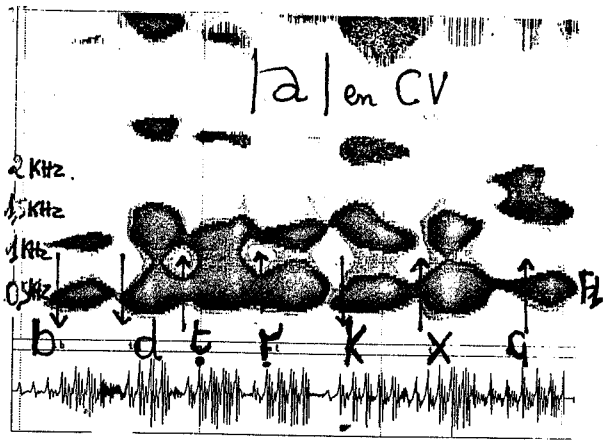


F1 et F2 des voyelles précédées des différents types de consonnes

□ les consonnes labiales et dentales
0 les consonnes emphatiques
✳ les consonnes pharyngales
△ les consonnes uvulaires

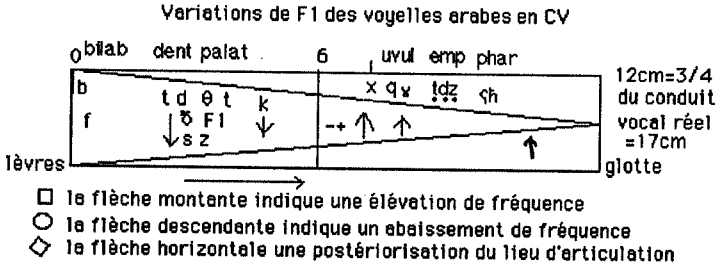
Les figures 1a et 1b montrent que
1) F1 des voyelles précédées des consonnes emphatiques, uvulaires et pharyngales est plus élevé [1,2,3].
2) F1 des voyelles précédées des consonnes labiales et dentales est nettement moins élevé.
3) La distance entre F1 et F2 des voyelles /a/, /a:/ ,/u/, /u:/ précédées d'une emphatique, d'une uvulaire et d'une pharyngale est moins grandepar rapport à celle des voyelles précédées des labiales et des dentales .

A titre d'exemple, ce spectogramme en LPC de la voyelle /a/ montre cette élévation de la zone de résonance de F1 au contact des types de consonnes précitées :



Résultats

L'élévation de F1 des voyelles précédées d'une emphatique, d'une uvulaire et d'une pharyngale semble être une constante par rapport aux voyelles précédées des labiales et des dentales. Cette élévation de F1 est due à une constriction pharyngale qui a été régulièrement constatée dans les radiofilms [1,3,4] et prédictible d'après les travaux sur la modélisation du conduit vocal [6,7,8] : une constriction au niveau du pharynx réduit le volume de cette cavité et relève par conséquent le premier formant en entraînant un changement de timbre. Voir ce schéma :



A noter que les consonnes emphatiques sont réalisées avec une double constriction en avant et en arrière du conduit vocal . Le modèle de G.Fant est construit sur une seule constriction. A titre d'exemple,

nous avons placé les consonnes emphatiques sur le tube acoustique entre les uvulaires et les pharyngales puisqu'elles sont corrélées à une constriction pharyngale avec un relèvement du premier formant de la voyelle adjacente.

Nous avons constaté le rôle important de F1 dans la synthèse et la perception à l'aide du modèle articulatoire de Maeda [9] en réduisant la fonction d'aire du premier tube qui correspond à la cavité postérieure. A titre d'exemple : A1 = 3 cm pour la voyelle antérieure [a] à F1 = 664 Hz et 2 cm pour la voyelle postérieure [ɑ] à F1 = 674 Hz.

Sur le plan phonologique, l'élévation de F1 est le corrélat acoustique proposé par N. Clements pour le trait "radical" [5] : selon ce phonologue, la constriction pharyngale corrélée avec le relèvement du premier formant est nécessaire pour définir l'ensemble des voyelles, en plus des traits de coronalité et de labialité.

Pour valider nos résultats préliminaires, nous avons appliqué une opération statistique appelée le Test T apparié sur les valeurs moyennes de F1 des voyelles des deux groupes de consonnes des trois locuteurs :

	Groupe 1 (X1)	Groupe 2 (X2)
consonnes	labiales	emphatiques
	dentales	uvulaires
		pharyngales

Les résultats de ce test T sont très significatifs . Ils confirment qu'il s'agit de deux classes de vocoïdes :
1) Pour lavoyelle /i/ (T[307]=4,5;p<0,05- dl =46 pour toutes les voyelles.
2) Pour la voyelle brève /a/ T= 6.
3) pour la voyelle brève /u/ T=5.
4) Pour la voyelle longue /i:/ T=4.
5) Pour la voyelle longue /a:/ T=6.
6) Pour la voyelle longue /u:/ T=3.

Nous présentons ici les valeurs moyennes de F1 et F2 des voyelles des deux groupes des trois locuteurs :

Figure 1c

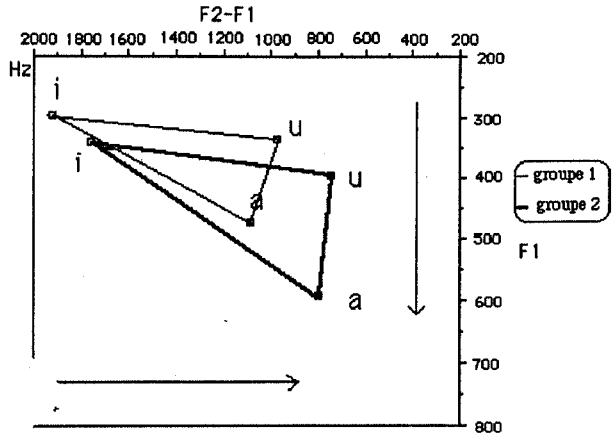
Tableau des valeurs moyennes de F1 et F2 des deux groupes de vocoïdes

	v	F1	F2
G1	i	296	2216
G2	i	341	2101
G1	a	473	1563
G2	a	592	1393
G1	u	343	1307
G2	u	396	1146
G1	i:	296	2351
G2	i:	352	2221
G1	a:	485	1720
G2	a:	578	1407
G1	u:	345	1065
G2	u:	376	1000

Nous avons visualisé sur la figure 1d les valeurs moyennes des formants des deux groupes de vocoïdes brèves sur le plan F1/F2.

Figure1d

Triangle acoustique des vocoïdes brèves



Le triangle acoustique montre que la distance entre F1 et F2 des vocoïdes du Groupe2 est moins grande que celle du groupe1. La flèche horizontale indique une postériorisation du lieu d'articulation pour ce groupe à F1 plus élevé (10). Cette postériorisation est corrélée avec un accroissement du degré d'aperture de la cavité buccale. La flèche verticale indique ce fait.

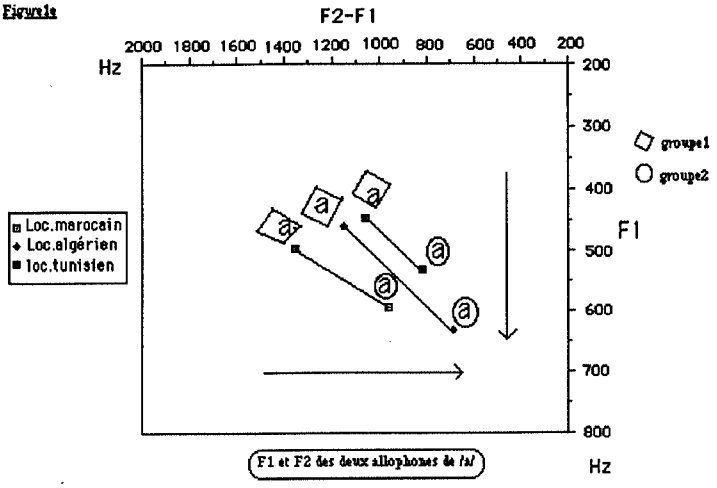
Discussion

La variabilité existe dans ces deux classes . Elle est liée aux facteurs suivants :

- la nature de la consonne uvulaire, pharyngale et emphatique.
- la nature du dialecte mais on ne peut pas généraliser à partir d'un seul locuteur.
- la durée de la voyelle : longue ou brève.

A titre d'exemple, nous présentons ici la figure 1e qui montre la moyenne des deux premiers formants de /a/ des trois locuteurs maghrébins sur le plan F1/F2 :

Figure1



F1 de la vocoïde (G1) du marocain est plus élevé que celui de la vocoïde correspondante chez le locuteur algérien et du tunisien . La distance entre F1 et F2 de la vocoïde (G2) est plus grande chez le locuteur algérien que chez le locuteur marocain ou tunisien . A notre avis, la constriction pharyngale est peut-être plus importante chez le locuteur algérien : elle est corrélée avec un plus grand degré d'aperture buccale. Le débit lent et soigné de ce locuteur semble induire ce phénomène.

Conclusion

L'analyse confirme l'élévation de F1 des voyelles précédées d'une emphatique, d'une uvulaire et d'une pharyngale par rapport à celle précédées des labiales et des dentales . Elle montre que cette élévation de F1 peut-être un indice acoustique utile pour une répartition binaire du système vocalique de l'Arabe standard en vue d'application dans la synthèse (11).

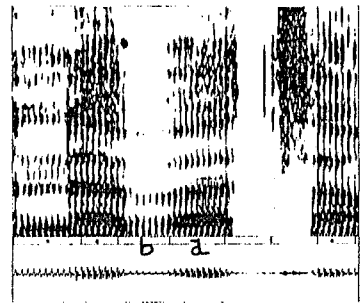
Remerciements

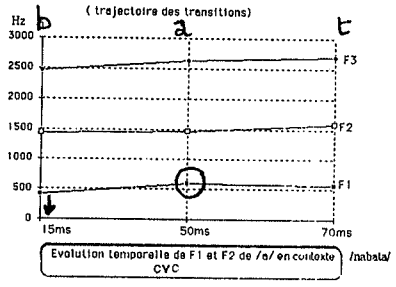
Je tiens à remercier Mmes J. Vaissiere, A. Rialland M. R. Gsell et M.J.Y.Dommergues de leurs conseils .

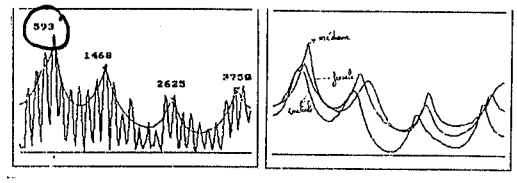
Bibliographie

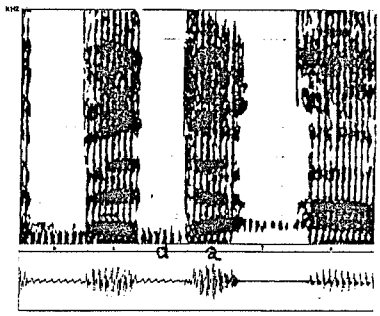
(1) Al Ani , S. H . (1970), *Arabic Phonology* , Mouton , The Hague.
(2) Kiel, N. (1987), "A phonetic Study of emphasis vowels in Egyptian Arabic ", *Working Papers* 13, Lund University
(3)Ghazeli, S. (1977), *Back Consonants Baking Articulation in Arabic* , Phd.Dissertation, University of Texas At Austin .
(4) Boff, F. (1983), "Contribution à l'étude expérimentale des consonnes d'arrière de L'arabe classique", Travaux de L'Institut de Strasbourg N°15.
(5) Clements, G.N. (à paraître), "Lieu d'articulation des voyelles et des consonnes : une théorie unifiée" dans *Architecture des représentations Phonologiques* , Ed. B. Laks et A. Rialland, Collection Sciences du langage.
(6) Delattre,P. " Pharyngeal Features in The Consonants of Arabic, German, French and American English, *Phonetica* 21, pp.129-155.
(7) Fant,G. (1973) *Speech Sounds And Features, Cambridge, England.*
(8) Santerre,L . (1985), " Corrélations entre les mouvements articulatoires et les variations formantiques", Travaux de l'Institut de Strasbourg, N°17, pp.389-400.
(9) Travaux dirigés de D.E.A sur le modèle articulatoire de Maeda à L'institut de Phonétique de Paris III, 1992.
(10) Ghazeli, S. (1981), "La coarticulation de l'emphase en arabe ", Arabica Tome2, pp.251-276.
[11] A paraître dans notre thèse" Nouveau Doctorat " à Paris III intitulée : *Contribution à l'étude formantique des voyelles de l'arabe standard.*

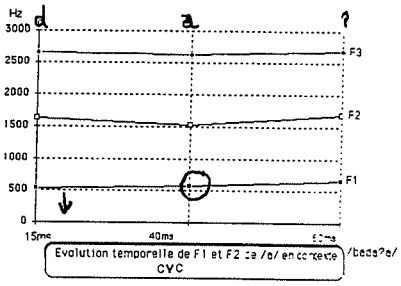
Nous présentons ici les spectogrammes des différentes réalisations de la voyelle /a/ en CV prononcée par notre locuteur marocain. Nous constatons l'élévation du F1 de cette voyelle précédées de la consonne emphatique [ṭ] , la consonne uvulaire [x] et de la consonne pharyngale [ʕ].

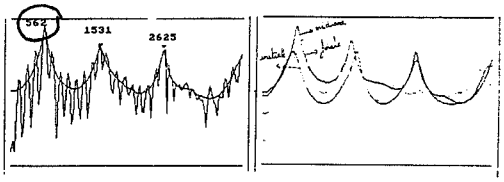


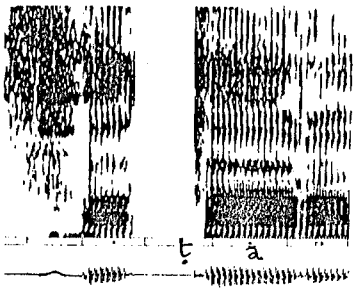


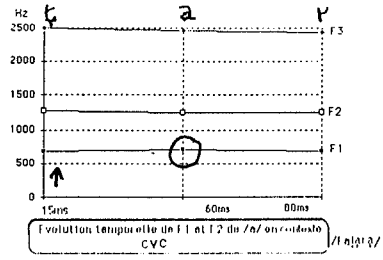


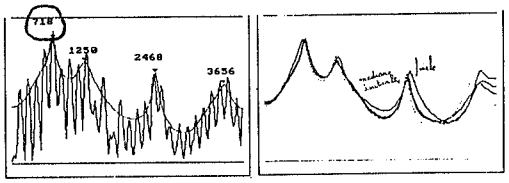

a a

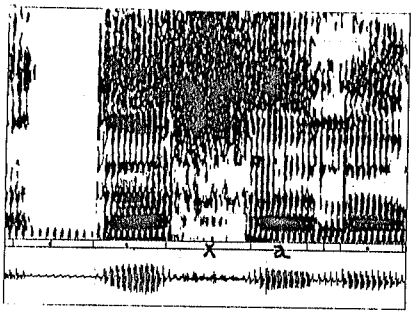

Hz
3000
2500
2000
1500
1000
500
0
d a ?
F3
F2
F1
15ms
40ms
Evolution temporelle de F1 et F2 de /a/ en contexte CVC
/bada?a/

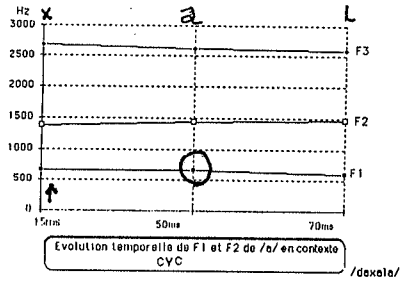

562
1531
2625

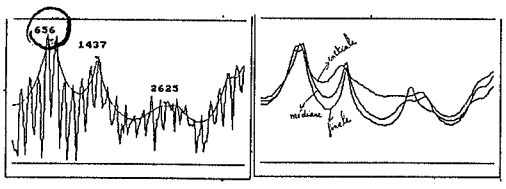

t a

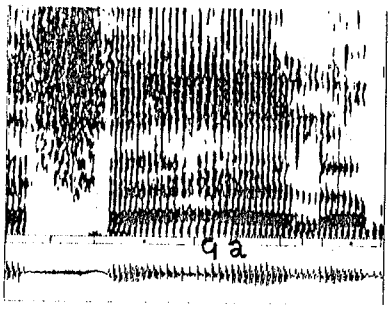

Hz
2500
2000
1500
1000
500
0
t a
F3
F2
F1
15ms
60ms
Evolution temporelle de F1 et F2 de /a/ en contexte CVC

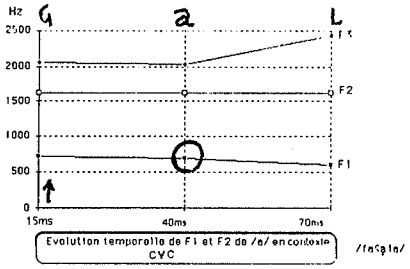

718
1259
2468
3656

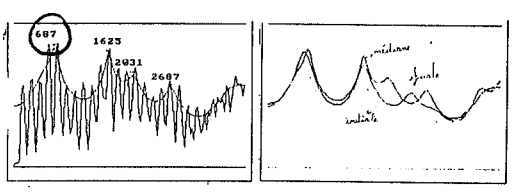

x a


Hz
3000
2500
2000
1500
1000
500
0
x a
F3
F2
F1
15ms
50ms
70ms
Evolution temporelle de F1 et F2 de /a/ en contexte CVC
/dexela/


636
1437
2625


ɡ a


Hz
2500
2000
1500
1000
500
0
ɡ a
F3
F2
F1
15ms
40ms
70ms
Evolution temporelle de F1 et F2 de /a/ en contexte CVC
/ɛɡa/


687
1625
2031
2687

19es J.E.P. - Bruxelles, 19 au 22 mai 1992.

Reconnaissance Automatique de la Parole: Modèles Stochastiques et/ou Modèles Connexionistes

Hervé Bourlard

Lernout & Hauspie Speech Products
Rozendaalstraat, 14
B-8900 Ieper, BELGIUM

Résumé

Dans ce papier, après un bref rappel des méthodes stochastiques utilisées pour la reconnaissance de la parole, et plus particulièrement des modèles de Markov cachés, nous considérons certains réseaux de neurones qui pourraient être utiles à cette tâche. Il semble cependant utopique d'imaginer des systèmes capables de reconnaître la parole, et plus particulièrement la parole continue, sur base de réseaux de neurones seulement. Nous nous efforçons donc de montrer ici comment les réseaux de neurones peuvent être utilisés pour améliorer les méthodes existantes. Une telle démarche requiert cependant une bonne connaissance et une bonne compréhension des différentes techniques, de façon à en optimiser leur interaction.

1 Introduction

Le caractère stochastique et séquentiel de la production de la parole rend sa reconnaissance automatique difficile. Ces dernières années, la théorie des modèles de Markov cachés, "Hidden Markov Models" (HMM) en anglais, a conduit à une bonne représentation de ces caractéristiques et a permis le développement de plusieurs systèmes de reconnaissance relativement performants. Cependant, malgré les améliorations constantes de ces algorithmes, il semble aujourd'hui qu'il sera difficile d'aboutir à des systèmes de reconnaissance pouvant traiter le language naturel, avec ses variations inter-locuteurs et son grand vocabulaire, par cette seule technique.

Alors que les approches de l'intelligence artificielle et de la coopération de différentes sources de connaissance sont conceptuellement plus satisfaisantes, elles n'ont, jusqu'à présent, pas conduit à des résultats acceptables en reconnaisance de la parole.

Récemment, de nouveaux modèles basés sur les réseaux de neurones ont fait leur apparition dans les problèmes de reconnaissance des formes et, naturellement, ont commencé à être appliqués au problème de la reconnaissance de la parole. Présentant plusieurs avantages qui seront discutés dans ce papier, ces modèles souffrent cependant de faiblesses plus particulièrement liées au traitement de signaux séquentiels. Nous nous efforcerons de démontrer que la solution optimale se situe probablement dans une approche modulaire où, partant d'un système existant, certains des modules les plus faibles sont remplacés par d'autres, plus performants, basés sur les réseaux de neurones. La reconnaissance de la parole se prête assez bien à ce genre de développement étant donné que ce problème, comme tous les problèmes difficiles, doit être décomposé en sous-problèmes, par exemple: bonne représentation des caractéristiques du signal, bonne définition des unités acoustiques utilisées (phonèmes, triphones, syllabes, mots, ...), bon algorithme de comparaison des formes (tenant compte des déformations temporelles), bonne interaction avec les connaissances syntaxiques, sémantiques et pragmatiques. L'avantage supplémentaire d'une telle approche modulaire est de conduire pas-à-pas, et par améliorations successives, vers une approche globale utilisant le même formalisme.

2 Modèles stochastiques

Selon le formalisme des modèles de Markov cachés (HMM), le signal de parole est supposé être produit par un automate stochastique fini construit à partir d'un ensemble d'états $\mathcal{Q} = \{q_1, q_2, \ldots, q_K\}$ et régi par des lois statistiques. Dans ce cas, chaque unité de parole (par exemple, chaque phonème ou chaque mot) est représenté par un modèle particulier. Le critère utilisé pour l'entraînement des paramètres et pour la reconnaissance est alors basé sur le maximum de vraisemblance, en anglais Maximum Likelihood Estimate (MLE), $P(X|M)$ représentant la probabilité qu'une séquence de vecteurs acoustiques $X = \{x_1, x_2, \ldots, x_N\}$ soit générée par un modèle

de Markov M. Pendant la phase d'entraînement, les paramètres des modèles sont estimés de façon à maximiser $P(X|M)$ où M est le modèle de Markov associé à la phrase d'entraînement qui a produit X et qui est simplement obtenu par la concaténation des modèles de Markov élémentaires correspondant à la séquence des unités de parole dans la phrase considérée (supposée connue lors de l'entraînement). Lors de la reconnaissance, on recherche la (meilleure) séquence de modèles de Markov M (et donc la meilleure séquence d'unités de parole) qui maximise $P(X|M)$.

Après avoir choisi a priori la topologie des modèles et avoir supposé que ces modèles de Markov sont d'ordre 1 (c'est-à-dire que l'influence des états précédents est limitée au dernier état rencontré), on peut montrer que $P(X|M)$ s'exprime en fonction de probabilités locales du type $p(x_t, q_k^t|q_\ell^{t-1})$ représentant la probabilité d'observer, à l'instant t, le vecteur x_t sur l'état q_k, alors que x_{t-1} avait été observé sur l'état q_ℓ. Afin de réduire le nombre de paramètres, ces probabilités locales sont souvent décomposées en un produit de probabilités d'emission $p(x_t|q_k)$ et de probabilités de transition $p(q_k|q_\ell)$.

Cette approche nous fournit une formulation élégante du problème de la reconnaissance de la parole et bénéficie d'algorithmes très efficaces pour la reconnaissance (algorithme de Viterbi, basé sur la programmation dynamique) et pour l'entraînement automatique (algorithme "Forward-Backward" de Baum-Welch) semi-supervisé (il suffit de connaître la séquence des unités de parole contenues dans les phrases d'entraînement, et il n'est pas nécessaire d'en avoir la segmentation).

Les hypothèses qui rendent l'optimisation de ces modèles possible limitent néanmoins leur généralité et sont à l'origine de certaines de leurs faiblesses, à savoir:

- Faible pouvoir discriminant de ces modèles qui, lors de l'entraînement, maximisent le critère de vraisemblance $P(X_j|M_j)$, si M_j est le modèle associé à X_j, mais ne minimisent pas la vraisemblance des classes rivales, à savoir $P(X_j|M_k)$, $\forall k \neq j$. L'utilisation de la probabilité a posteriori $P(M|X)$ résout ce problème mais semble difficile à formuler proprement dans le formalisme des modèles de Markov.

- Choix a priori de la topologie des modèles et des distributions statistiques régissant les probabilités d'émission (souvent supposées gaussiennes ou multi-gaussiennes).

- Restriction aux modèles de Markov d'ordre 1.

- L'information contextuelle des vecteurs acoustiques n'est pas prise en compte (on a négligé le corrélation entre vecteurs acoustiques).

- Le formalisme très particulier des modèles de Markov rend leur intégration avec d'autres sources de connaissance (comme la syntaxe, la sémantique et la pragmatique) difficile.

Pour un bon aperçu général des modèles de Markov, voir [Jelinek, 1976; Lee, 1989; Rabiner, 1989].

3 Réseaux de neurones

Etant donné les limitations des modèles de Markov, plusieurs groupes de recherches ont entrepris l'étude des réseaux de neurones, et plus particulièrement des perceptrons multicouches, en anglais "Mulilayer Perceptrons" (MLP), pour la reconnaissance automatique de la parole.

Nous avons cependant jugé bon de commencer cette section par certaines affirmations erronées (mais couramment rencontrées) concernant les réseaux de neurones.

Affirmation: Nous comprenons la parole grâce à un réseau biologique de neurones (notre cerveau!); en conséquence, les réseaux de neurones, qui sont un modèle de ce réseau biologique, devraient être mieux adaptés à la reconnaissance de la parole que les modèles purement mathématiques.

Bien que cette assertion soit fondamentalement valable, il ne faut pas oublier que les réseaux de neurones utilisés actuellement ne sont que des modèles extrêmement grossiers des réseaux biologiques. Même si ces modèles étaient précis, il faudrait encore savoir comment traiter et faire interagir un très grand nombre de cellules élémentaires tout en respectant un certain nombre de contraintes liées aux possibilités d'implémentation et imposées par les ingénieurs.

Affirmation: Les réseaux de neurones sont des systèmes adaptatifs. Par conséquent, ils sont capables d'apprendre et de généraliser certaines informations.

Avant de commencer à travailler avec les réseaux de neurones, il est toujours bon de se rappeler la longue histoire des classificateurs entraînables, parmi lesquels nous retrouvons d'ailleurs toutes les approches statistiques et les modèles de Markov. La plupart

de ces systèmes peuvent également être entraînés de façon adaptative, et souvent beaucoup plus rapidement que les algorithmes d'apprentissage liés aux réseaux de neurones.

Affirmation: Les systèmes classiques de reconnaisance de formes nécessitent souvent l'introduction de connaissances et d'hypothèses spécifiques à l'application traitée, ce qui limite leurs possibilités et conduit à des performances sous-optimales.

Il est vrai que les systèmes adaptatifs comme les réseaux de neurones fournissent un bon mécanisme pour l'apprentissage paramétrique sans nécessiter trop d'hypothèses simplificatrices. Cependant, pour des problèmes suffisamment larges, ces systèmes devront, en général, être contraints également, et des informations spécifiques à l'application traitée sont souvent la meilleure façon de définir ces contraintes. De plus, des choix arbitraires tels que le nombre de couches et d'unités cachées, ainsi que les paramètres des algorithmes d'entraînement utilisés, sont toujours sous-jacents à l'utilisation de ces modèles.

Affirmation: Une affirmation apparentée à la précédente consiste à dire que les méthodes de classification traditionelles nécessitent la sélection (arbitraire) de caractéristiques préalablement extraites du signal original, alors que les réseaux de neurones pourraient faire cela automatiquement et, par conséquent, éliminer cette étape sous-optimale.

De nouveau, il est vrai que la détermination automatique des paramètres optimaux caractérisant le signal est hautement souhaitable. Il est également vrai que, dans certains cas (comme le système développé chez AT&T pour la reconnaisance automatique de codes postaux), il semble possible d'extraire les caractéristiques pertinentes automatiquement à partir du signal brut. Cependant, dans la plupart des cas, les données brutes contiennent toujours beaucoup trop d'informations et un simple pré-traitement permet souvent d'améliorer les performances des systèmes de reconnaissance. Evidemment, il n'est jamais facile de savoir ce qui doit être pre-déterminé et ce qui doit être appris, mais il semble certain qu'une bonne sélection d'un ensemble (éventuellement élargi) de caractéristiques sera toujours requis dans les systèmes utilisant les réseaux de neurones.

Si ces affirmations sont effectivement erronées, on peut alors se demander ce qu'il reste du potentiel des réseaux de neurones pour la reconnaissance de la parole. Comme les modèles de Markov, les réseaux de neurones ont l'avantage de pouvoir être entraînés. Ils ont en outre plusieurs propriétés attrayantes, à savoir:

- Leur entraînement est basé sur des critères discriminants.
- Lorsqu'ils sont utilisés en mode de classification, ces réseaux peuvent estimer les probabilités a posteriori sans devoir recourir à des hypothèses concernant les distributions statistiques des données.
- Les réseaux de neurones peuvent facilement incorporer différents types de représentation et en trouver les combinaisons optimales sans nécessiter d'hypothèses d'indépendence.
- Grâce à leurs propriétés d'interpolation, ces réseaux sont capables de bonnes classifications statistiques sur des espaces sous-échantillonnés [Niles et al., 1989] sans nécessiter trop d'hypothèses simplificatrices.
- Bien que nous ayons suggéré plus haut que l'élimination complète des heuristiques en reconnaisance de la parole soit difficile, il est néanmoins raisonnable de penser que certaines hypothèses arbitraires peuvent être remplacées par des procédures d'apprentissage adaptatif.
- Les réseaux de neurones sont caractérisés par des architectures massivement parallèles, ce qui les rend particulièrement intéressants pour les implémentations "hardware" rapides.

Malheureusement, ces réseaux de neurones ont également des faiblesses lorsqu'on essaye de les utiliser pour la reconnaissance de la parole; en particulier, ceux-ci sont mal adaptés au traitement des informations séquentielles.

4 Réseaux de neurones et séquences temporelles

Nous considérons ici le problème de l'apprentissage de séquences temporelles par réseaux de neurones. Il est bon de distinguer trois tâches différentes:

- Reconnaissance de séquences- Dans ce cas on veut apprendre au réseau à produire une sortie particulière lorsqu'une séquence a été présentée à l'entrée. Il n'est pas nécessaire

de classer chaque composant de la séquence indépendamment. Ceci est la tâche typique de la reconnaissance de la parole, pour autant que la séquence d'entrée ne contienne qu'une des unités de parole (mots isolés ou phonèmes isolés) avec lesquelles les états de sortie sont associés.

- Reproduction de séquences- Dans ce cas, le réseau doit être capable de compléter une séquence particulière seulement sur base d'une partie de celle-ci. Ceci est donc une généralisation de l'auto-association à des entrées dynamiques. Ceci peut être utilisé par exemple pour modéliser certains processus auto-régressifs ou pour prédire la suite d'un série chronologique.

- Association temporelle- Dans ce cas, une séquence particulière doit être produite à la sortie du réseau en réponse à une séquence d'entrée spécifique. Les séquences à l'entrée et à la sortie sont alors généralement différentes: ceci est donc une forme d'hétéro-association généralisée aux formes dynamiques. Par exemple, ceci pourrait être utilisé, du moins en théorie, pour la reconnaissance de parole continue où la séquence d'entrée est une suite de vecteurs acoustiques et où la séquence de sortie serait une suite de mots ou de phonèmes. Cependant, dans ce cas, les séquences à l'entrée et à la sortie seront asynchrones (ce qui peut poser un sérieux problème) étant donné que chaque composant de la séquence de sortie (par exemple, un mot) doit être associé à une sous-séquence de l'entrée.

Nous parcourons maintenant brièvement les différentes architectures qui ont été proposées pour résoudre l'une ou plusieurs de ces tâches.

4.1 Registres à décalage

La façon la plus simple de reconnaître une séquence temporelle est de transformer celle-ci en une représentation spatiale à l'entrée d'un réseau de neurones contenant un registre à décalage. Cette méthode a été utilisée avec succès pour reconnaître des mots ou des phonèmes isolés [Landauer et al., 1987; Peeling & Moore, 1988].

Cette approche souffre cependant de plusieurs inconvénients, à savoir:

- Le registre doit être suffisamment large que pour pouvoir contenir toute la séquence, ce qui augmente le nombre de paramètres et donc le nombre d'exemples nécessaires à l'entraînement.

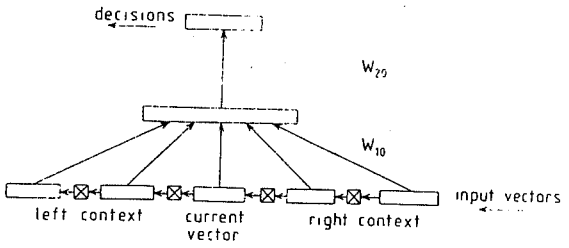


Figure 1: MLP avec registre à décalage.

- Le réseau n'est pas automatiquement invariant à la distorsion et au décalage, et il est souvent nécessaire de présenter un très grand nombre d'exemples pour chaque classe de sortie et pour toutes les positions possibles dans le registre.

- Cette approche ne semble pas adaptée à la reconnaissance de parole continue, c'est-à-dire de séquences dans lesquelles se trouvent plus d'une classe de sortie.

4.2 Réseaux récurrents

La formule idéale de réseau pour la reconnaissance de séquences est sans doute les réseaux récurrents. Dans ce cas, chaque état du réseau est récurrent et la sortie à l'instant $t-1$ est utilisée pour calculer l'activation à l'instant t. Dans le cas de l'association temporelle, on essaie de produire une séquence de sortie pour chaque séquence spécifique à l'entrée. La sortie désirée utilisée pour l'entraînement est alors fixée à des valeurs bien précises à certains instants particuliers de la séquence (par exemple, la valeur du mot à la fin de chaque mot). Lorsque la longueur maximale T des séquences à reconnaître n'est pas trop grande, on peut utiliser l'algorithme d'entraînement standard des perceptrons multicouches en "dépliant" le réseau dans le temps, comme suggéré dans [Minsky & Papert, 1969; Rumelhart et al., 1986]. Cet algorithme est connu sous le nom de "rétro-propagation dans le temps". Différentes versions de cet algorithme ont été proposées et testées sur des tâches difficiles mais limitées à quelques mots ou phonèmes isolés [Watrous & Shastri, 1987]. Cependant, pour des séquences plus longues, cette approche devient rapidement impraticable. Dans [Kuhn et al., 1990], il est montré qu'il est possible d'éviter la rétro-propagation dans

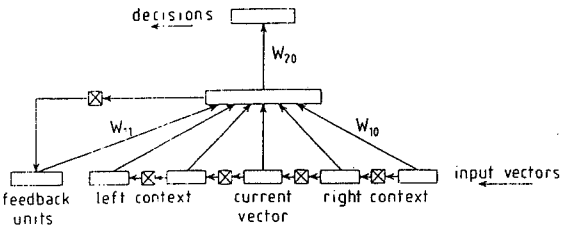


Figure 2: MLP avec rétroaction des unités cachées.

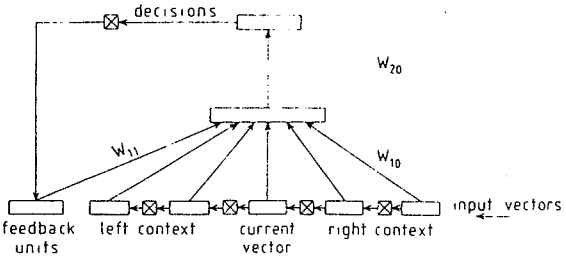


Figure 3: MLP avec rétroaction des unités de sortie.

le temps au prix de beaucoup plus de dérivées partielles à garder lors de la propagation des activations ("forward propagation") dans le réseau.

4.3 Réseaux à rétroaction

Une autre façon assez populaire de traiter les sequences temporelles est d'utiliser des réseaux partiellement récurrents avec simple rétroaction des unités cachées ou des unités de sortie vers l'entrée. Alors que ceci ne complique pas significativement le réseau, il lui permet de mémoriser le passé récent.

Dans [Elman, 1988], l'activation des unités cachées à l'instant $t-1$ est reproduite à l'entrée dans des unités de rétroaction ("feedback units") qui viennent s'ajouter aux observations relatives à l'instant t. Il a été montré que ce type de réseau était capable de reconnaître des séquences et aussi de compléter de courtes séquences connues. Dans [Cleeremans et al., 1989], il a été montré que ce type de réseau était également capable de représenter un automate fini (par exemple, pour la modélisation de grammaires).

Un autre type de réseau à rétroaction a également été proposé dans [Jordan, 1986]; dans ce cas, l'activation des unités de sortie est renvoyée vers l'entrée pour compléter les observations relatives à l'instant t. Initialement, cette approche était utilisée pour la production de séquences, dans laquelle chaque entrée fixe générait une séquence particulière à la sortie. Ce type de réseau peut également être utilisé pour la classification de séquences. De plus, dans [Bourlard & Wellekens, 1990] on montrait que, dans ce cas, il était possible d'interpréter les valeurs aux sorties comme des probabilités à posteriori (c'est-à-dire les probabilities des classes de sorties conditionnées par les entrées) qui pouvaient alors être utilisées comme probabilités d'emission dans des modèles de Markov.

Il est intéressant de noter que ces réseaux récurrents peuvent également être approchés par des réseaux à registre. Dans ce cas, nous avons la solution proposée dans [Lang et al., 1990].

4.4 Discussions

Tous ces modèles ont conduit à de bonnes performances sur des unités de parole isolées, pour autant que celles-ci ne soient pas trop longues. Grâce à leur dynamique implicite, ils sont capables d'intégrer des informations temporelles et, dans certaines limites, de traiter les déformations temporelles. Cependant, l'entraînement et la reconnaissance sur des séquences contenant plusieurs unités de parole est toujours impossible en utilisant uniquement des réseaux de neurones. En ce qui concerne l'entraînement, il y a également le problème de définir les sorties désirées, ce qui est difficile si les données ne sont pas préalablement segmentées. Même si la segmentation est connue, la valeur à imposer aux sorties pour l'intérieur des segments n'est toujours pas claire, et on se contente généralement de fonctions linéaires croissant de 0 à 1 du début du segment à sa fin. Pour la reconnaissance, on ne sait toujours pas comment traiter les sorties lorsque la séquence d'entrée contient plusieurs unités de parole. Dans ce cas, il est vrai que les valeurs des unités de sortie passeront chronologiquement par des minima et des maxima correspondant respectivement à l'absence ou à la présence d'unités de parole mais il est également clair que ces informations devront être filtrées pour conduire à une reconnaissance globale optimale. En fait, c'est ce que fait la programmation dynamique dans le formalisme des modèles de Markov.

5 Simulation de HMM par réseaux de neurones

Etant donné le succès des modèles de Markov, plusieurs groupe de recherche se sont efforcés à représenter des modèles de Markov selon le formalisme des réseaux de neurones. Deux résultats intéressants sont à mentionner ici.

Dans [Lippmann & Gold, 1987], il est montré qu'un réseau de neurones particulier, appelé réseau Viterbi, peut parfaitement simuler les fonctions de l'algorithme de reconnaissance Viterbi des modèles de Markov.

Dans [Kehagias, 1989; Bridle, 1990], on montre également qu'il est possible de reformuler parfaitement l'algorithme d'entraînement des modèles de Markov ("Forward-Backward") en terme d'une variante de l'algorithme de retro-propagation dans le temps appliqué à une forme particulière de réseaux de neurones récurrents.

Ces résultats sont très utiles pour l'interprétation des différents algorithmes dans un formalisme commun. Il est cependant difficile d'en tirer avantage de façon à améliorer les méthodes existantes.

6 Approches Hybrides

Dans les approches hybrides, l'idée est d'essayer de combiner proprement les approches standards des modèles de Markov avec certaines techniques relatives aux réseaux de neurones, afin de tirer parti des avantages respectifs de chacune de ces approches. Dans ce but, il a été démontré dans [Bourlard & Wellekens, 1990] que, lorsque des réseaux de neurones, et en particulier des perceptrons multicouches (MLP), sont entraînés de façon à minimiser un critère de moindres carrés ou un critère d'entropie, les valeurs optimales obtenues à la sortie des ces réseaux peuvent être parfaitement interprétées en terme de probabilités a posteriori (également appelées probabilités Bayesiennes) des classes de sorties conditionnées par les entrées. En d'autres mots, si q_k représente une sortie particulière du réseau et x_n une entrée particulière à l'instant n, l'entraînement du réseau de neurones dans le cas d'un critère de moindres carrés minimisera la fonction

$$E = \sum_{n=1}^{N} \sum_{k=1}^{K} [g_k(x_n) - d_k(x_n)]^2 \quad ,$$

dans laquelle $g_k(x_n)$ et $d_k(x_n)$ représentent respectivement l'activation observée et la valeur idéale de la sortie associée à q_k lorsque l'on présente x_n à l'entrée. En mode de classification (ce qui est le cas en reconnaissance de la parole), nous avons $d_k(x_n) = \delta_{k\ell}$ si on sait que $x_n \in q_\ell$.

Dans [Bourlard & Wellekens, 1990], il a été démontré que si le réseau contient suffisamment d'unités cachées et si l'entraînement ne converge pas vers un minimum local, les sorties optimales obtenues étaient alors données par:

$$g_k^{opt} = p(q_k|x_n) \quad .$$

Cette conclusion peut se généraliser facilement à d'autres types de réseaux de neurones. Si un réseau comme celui représenté à la figure 1 est utilisé, dans lequel l'entrée ne contient pas seulement l'observation courante x_n mais également son contexte de gauche et de droite, en introduisant un fenêtre temporelle de largeur $2c+1$ contenant $\{x_{n-c}, \ldots, x_n, \ldots, x_{n+1}\}$, le réseau estimera des probabilités du type $p(q_k|x_{n-c}, \ldots, x_n, \ldots, x_{n+1})$ qui peuvent également être utilisées dans des modèles de Markov, permettant ainsi d'éviter l'hypothèse standard consistant à négliger la corrélation entre observations successives.

Une autre possibilité est d'utiliser le réseau présenté à la figure 3. Dans ce cas, les valeurs générées sur les unités de sortie seront des estimateurs de $p(q_k|q_\ell^-, x_{n-c}, \ldots, x_n, \ldots, x_{n+1})$, probabilités qui tiennent maintenant compte non seulement de la corrélation des observations mais également de la classe q_ℓ^- observée à l'instant précédent.

Toutes ces probabilités peuvent s'intégrer assez facilement dans le formalisme des modèles de Markov (en fait, à la condition de modifier plusieurs éléments dans ce schéma de base [Bourlard & Morgan, 1990, 1991]), cumulant ainsi les avantages respectifs des deux approches, à savoir:

- Pour les modèles de Markov: bonne modélisation du caractère séquentiel de la parole et algorithmes efficaces pour le reconnaissance et pour l'entraînement (pour lequel on ne doit pas connaître au préalable la segmentation des phrases d'entraînement).
- Pour les réseaux de neurones: pas d'hypothèses concernant les distributions statistiques et possibilité de tenir compte de la corrélation entre les observations successives.

Récemment, il a été démontré expérimentalement que cette approche permettait effectivement d'améliorer significativement les résultats de reconnaissance actuellement disponibles, et cela même sur des tâches complexes comme la reconnaissance de la parole continue, grand vocabulaire (1000 mots) et indépendante du locuteur [Bourlard & Morgan, 1991; Bourlard & Morgan, 1992a].

7 Décomposition de larges réseaux de neurones

Un autre avantage de cette interprétation statistique des sorties est de pouvoir définir proprement une nouvelle méthode pour traiter les réseaux de neurones caractérisés par un très grand nombre de classes de sortie. En effet, dans les expériences dont il est fait mention dans les sections précédentes, chaque sortie du réseau était associée a une unité phonétique, ce qui limitait automatiquement le nombre de sorties à environ 60. Or, il est bien connu que, pour obtenir une reconnaissance robuste de la parole continue avec grands vocabulaires, il est utile de tenir compte de l'information contextuelle et de modéliser alors les phonèmes différemment en fonction de leur contexte phonétique, conduisant alors à la notion de diphones ou de triphones [Lee, 1990]. Dans ce cas, si q_k représente un phonème particulier parmi un ensemble de K phonèmes, il est alors nécessaire, par exemple pour les triphones, d'estimer des probabilités du type $p(q_\ell, q_k, q_r|x_n)$, où q_ℓ et q_r représentent respectivement le phonème à gauche et à droite de q_k. L'application directe de l'approche développée ci-dessus conduirait alors à un réseau de neurones contenant $K \times K \times K$ unités de sortie, si K représente le nombre de phonèmes de base, rendant ainsi son entraînement et son utilisation absolument impossible. Grâce à l'interprétation statistique exposée précédemment, il est possible de résoudre ce problème simplement en observant qu' en appliquant des règles élémentaires de statistique, nous pouvons écrire:

$$p(q_\ell, q_k, q_r|x_n) = p(q_\ell|q_k, q_r, x_n).p(q_r|q_k, x_n).p(q_k|x_n)$$

ce qui nous donne une méthode strictement correcte (et sans hypothèses sous-jacentes) pour factoriser un gros réseau de neurones en plusieurs réseaux plus simples. En effet, le réseau estimant la probabilité à gauche de l'égalité (et nécessitant $K \times K \times K$ unités de sorties) peut donc être décomposé en trois réseaux de neurones estimant respectivement un des facteurs à droite de l'égalité et ne contenant chacun que K unités de sortie. Un de ces réseaux (le troisième facteur) n'est rien d'autre que le réseau utilisé précédemment et representé à la figure 1. Les deux premiers facteurs peuvent être estimés par des réseaux similaires dans lesquels les sorties représentent respectivement les contextes phonétiques à gauche et à droite de q_k et où les entrées sont étendues aux classes présentes dans les conditionelles. Par exemple, le réseau estimant $p(q_\ell|q_k, q_r, x_n)$ est représenté à la figure 4. Des résultats préliminaires ont montré que cette approche permettait en effet d'estimer des probabilités dépendant du contexte phonétique [Bourlard & Morgan, 1992b; Morgan & Bourlard, 1992].

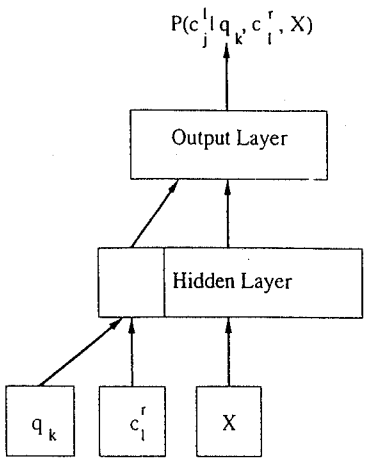


Figure 4: Réseau de neurone pour l'estimation de probabilités dependantes du contexte

8 Conclusions

Dans ce papier, nous avons tenté de montrer comment les réseaux de neurones peuvent être utilisés pour améliorer effectivement les systèmes existants de reconnaissance de la parole. Alors que ces réseaux de neurones ont des propriétés attrayantes et sont capables de traiter, dans certaines limites, les signaux séquentiels, il n'est cependant pas clair qu'il soit possible de reconnaître la parole continue uniquement avec ceux-ci. Grâce à une interprétation statistique de ces réseaux, nous avons montré qu'il était cependant possible de les utiliser conjointement avec les techniques des modèles de Markov afin de tirer parti des avantages respectifs des deux approches.

Lorsque ces réseaux de neurones sont utilisés en mode de classification, cette interprétation statistique conduit également à une méthode permettant de manipuler des réseaux contenant un très grand nombre d'unités de sortie sans nécessiter d'hypothèses simplificatrices.

References

[1] Bourlard, H., & Wellekens, C.J. (1990). Links Between Markov Models and Multilayer Perceptrons, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 12, no. 12, pp. 1167-1178.

[2] Bourlard, H., & Morgan, N. (1990). A Continuous Speech Recognition System Embedding MLP into HMM, D. S. Touretzky (ed.), *Advances in Neural Information Processing Systems 2*, pp. 186-193, San Mateo, CA: Morgan Kaufmann.

[3] Bourlard, H., & Morgan, N. (1991). Connectionist Approaches to the Use of Markov Models for Speech Recognition, R.P. Lippmann, J.E. Moody, & D.S. Touretzky (eds.), *Advances in Neural Information Processing Systems 3*, pp. 213-219, San Mateo, CA: Morgan Kaufmann.

[4] Bourlard, H., & Morgan, N. (1992a). Continuous Speech Recognition by Connectionist Statistical Methods, in preparation..

[5] Bourlard, H., & Morgan, N. (1992b). CDNN: A Context Dependent Neural Network for Continuous Speech Recognition, to be published in *IEEE Proc. Intl. Conf. on Acoustics, Speech, and Signal Processing*, San Francisco, CA, 1992.

[6] Bridle, J.S. (1990). Alpha-Nets: A Recurrent "Neural" Network Architecture with a Hidden Markov Model Interpretation, *Speech Communication*, vol. 9, no. 1, pp. 83-92.

[7] Cleeremans, A., Servan-Schreiber, D., & McClelland, J.L. (1989). Finite State Automata and Simple Recurrent Networks, *Neural Computation*, vol. 1, pp.372-381.

[8] Elman, J.L. (1988). Finding Structure in Time, *CRL Tech. Report 8801*, University of California at San Diego, Center for Research in Language, also in *Cognitive Science*, vol. 14, pp. 179-211.

[9] Jelinek, F. (1976). Continuous Recognition by Statistical Methods, *Proceedings IEEE*, vol. 64, no.4, pp. 532-555.

[10] Jordan, M. (1986). Attractor Dynamics and Parallelism in a Connectionist Sequential Machine, *Proc. of the Eighth Annual Conference of the Cognitive Science Society* (Amherst 1986), pp. 531-546, Hillsdale: Erlbaum.

[11] Kehagias, A. (1989). Optimal Control for Training: The Missing Link Between Hidden Markov Models and Connectionist Networks, Division of Applied Mathematics Technical Report, Brown University, Providence, RI.

[12] Kuhn, G., Watrous, R.L., & Ladendorf, B. (1990). Connected Recognition with a Recurrent Network, *Speech Communication*, vol. 9, no. 1, pp. 41-48.

[13] Landauer, T.K., Kamm, C.A., & Singhal, S. (1987). Learning a Minimally Structured Back Propagation Network to Recognize Speech, *Proc. of the Ninth Annual Conf. of the Cognitive Science Society*, pp. 531-536.

[14] Lang, K.J., Waibel, A.H., & Hinton, G.E. (1990). A Time-Delay Neural Network Architecture for Isolated Word Recognition, Neural Networks, vol. 3, no. 1, pp. 23-43.

[15] Lee, K.F. (1989). *Automatic Speech Recognition - The Development of the Sphinx System*, Kluwer Academic, Norwell Mass.

[16] Lee, K.F. (1990). Context-Dependent Phonetic Hidden Markov Models for Speaker-Independent Continuous Speech Recognition, *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 38, no. 4, pp. 599-609.

[17] Lippmann, R.P. & Gold, B. (1987). Neural Classifiers Useful for Speech Recognition, *IEEE Proc. of the First Intl. Conf. on Neural Networks*, vol. IV, pp. 417-422.

[18] Morgan, N., & Bourlard, H. (1992). Factoring Networks by a Statistical Method, accepted fon publication in *Neural Computation*.

[19] Minsky, M. & Papert, S. (1969). *Perceptrons*, Cambridge, MA: MIT Press.

[20] Niles, L.T., Siverman, H., Tajcham, G., & Bush, M. (1989). How Limited Training Data Can Allow a Neural Network Classifier to Outperform an "Optimal Statistical Classifier", *IEEE Proc. Intl. Conf. on Acoustic, Speech, and Signal Processing*, pp. 17-20, Glasgow, Scotland.

[21] Peeling, S.M. & Moore, R.K. (1988). Isolated Digit Recognition Experiments Using the Multi-Layer Perceptron, *Speech Communication*, vol. 7, pp. 403-409.

[22] Rabiner, L.R. (1989). A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition, *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257-285.

[23] Rumelhart, D.E., Hinton, G.E. & Williams, R.J. (1986). *Parallel Distributed Processing. Exploration of the Microstructure of Cognition. vol. 1: Foundations*, Ed. D.E.Rumelhart & J.L.McClelland, MIT Press.

[24] Watrous, R.L. & Shastri, L. (1987). Learning Phonetic Features Using Connectionist Networks: An Experiment in Speech Recognition, *First International Conference on Neural Networks*, vol.2, pp.619-627, San Diego.

19es J.E.P. - Bruxelles, 19 au 22 mai 1992.

FORMES PROSODIQUES ET FOCALISATION SEMANTIQUE

CLAIRE GERARD et DELPHINE DAHAN

LABORATOIRE DE PSYCHOLOGIE EXPERIMENTALE, URA CNRS 316,
UNIVERSITE RENE DESCARTES

Résumé

Quelles sont les modifications prosodiques de la parole lors d'une focalisation sur des mots-cibles dans la lecture à voix haute ? La première étude précise l'amplitude et la localisation des variations temporelles réalisées par deux locuteurs lors de la lecture de 32 textes. La deuxième examine les variations temporelles, intensives, mélodiques, réalisées par 8 locuteurs lors de l'énonciation de phrases isolées. La nature syntactico-sémantique des cibles ne semble pas jouer sur les variations temporelles dans la lecture continue de textes, mais change les contours mélodiques dans l'énonciation de phrases.

La communication parlée implique que locuteur et auditeur partagent des connaissances implicites non seulement sur les structures syntaxiques et sémantiques du langage, mais aussi sur les structures prosodiques de la parole. Pour exprimer telle ou telle volonté expressive, le locuteur réalise des variations prosodiques qui sont identifiées par l'auditeur par référence à des représentations mentales stockées en mémoire (Sorin, 1989). Nous étudions l'insistance ou emphase sur un mot, et cherchons à mettre en évidence la production de formes, temporelles ou mélodiques, correspondant à la représentation mentale de cette focalisation.

L'accent d'emphase a une fonction rhétorique. Est-il caractérisé, comme l'accent "primaire", par des variations des trois paramètres prosodiques ? Une syllabe accentuée voit en effet la durée de sa voyelle s'allonger, sa hauteur tonale et son pic d'amplitude s'élever (Cutler, 1976 ; Huggins, 1978 ; Nooteboom, Brokx & de Rooij, 1978). La focalisation sémantique a pour fonction de souligner le choix d'un mot parmi les autres mots possibles (axe paradigmatique), mais aussi de marquer le contraste avec les autres mots de l'énoncé (axe syntagmatique) qui sont désaccentués (Séguinot, 1977 ; Cooper, Eady & Mueller, 1985 ; Eady & Cooper, 1986 ; Cutler & Isard, 1980). Les deux recherches présentées ici visent à préciser les modifications prosodiques qu'entraîne la focalisation d'un mot sur ce mot et sur son entourage. La première recherche est axée sur les variations temporelles lors de la lecture à haute voix de textes, la deuxième sur l'étude pluri-paramètrique du mot focalisé lui-même dans la lecture ou la répétition d'une phrase.

EXPERIENCE 1.(*)

L'étude des **variations temporelles** a porté sur un corpus étendu (environ 3 heures 45 de parole) correspondant à la lecture par deux locuteurs de 16 textes, chacun présenté deux fois : 1 - sans focus (versions dites "monotones") ; 2 - avec 14 mots-cibles, signalés par un changement de typographie (versions dites "en relief"). Chaque texte se présentait visuellement sous forme d'une série de paragraphes distincts et chaque paragraphe était divisé en phrases et sous-groupes syntaxiques marqués par la ponctuation. Entre deux signes ponctués, un groupe de sens se présentait comme une unité sémantique, et les mots-cibles étaient insérés à l'intérieur de ces unités. La motivation première de cette recherche était de préciser les durées de pauses en lecture continue (Gérard, Dahan & Rigaut, 1991), mais les premières observations ont conduit à développer d'autres analyses dont voici les résultats.

1. LE RALENTISSEMENT DU DEBIT.

La mesure traditionnelle du débit de parole (en nombre de syllabes par seconde) permettait d'abord de constater que les deux locuteurs ralentissent leur débit dans les versions en relief. Le locuteur 1 lisait globalement plus vite que le locuteur 2, mais le pourcentage de ralentissement quand on passe de la version monotone à la version en relief était du même ordre pour les 2 locuteurs : proche de 10 %. Sur un texte de 231 mots en moyenne, où 6% des mots devaient être accentués (14 cibles/231 mots), on constate donc 10 % de ralentissement total par rapport à une lecture monotone. La question est maintenant de savoir où se situent exactement ces ralentissements. Pour chaque texte, la durée totale de lecture a été calculée, puis la durée totale pour les 14 cibles des pauses précédant les cibles (codées PAV = pauses avant), et des pauses les suivant (PAP = pauses après), ainsi que la durée d'énonciation de la dernière syllabe précédant les 14 cibles (SYLL) et de la voyelle au sein de cette syllabe (VOY), enfin la durée d'énonciation des cibles (DE). La somme des allongements observés à proximité des cibles (correspondant à la succession : SYLL+PAV+DE+PAP) a alors été rapportée à l'allongement total du texte. Pour le locuteur 1, le ralentissement de la parole était imputable pour 69 % à l'entourage immédiat des cibles (qui, rappelons-le, ne représentaient que 6 % des mots du texte), et pour le locuteur 2, cette proportion était de 51%. Les modifications temporelles liées à la focalisation n'ont donc pas été réparties exactement de la même façon par les deux locuteurs, le locuteur 1 les "concentrant" plus autour des cibles. Nous avons soumis chacun des indices PAV, PAP, DE, SYLL, à des analyses de variance pour vérifier que les allongements constatés étaient significatifs : ils le sont tous, pour les deux locuteurs. Le tableau 1 présente les valeurs moyennes obtenues pour les versions monotones (F) et en relief (F) en réunissant ces deux locuteurs.

Tableau 1: Durées moyennes (ms)

	SYLL	PAV	DE	PAP
F	243	95	767	322
F	319	224	1027	600

Puisque l'entourage immédiat de la cible n'expliquait pas la totalité du ralentissement, il nous a semblé intéressant de savoir "à partir de quand et jusqu'à quand" le locuteur ralentissait son élocution. Nous avons donc exploré la chaîne parlée sur des distances plus lointaines des cibles. Cette analyse porte sur un corpus plus restreint: 23 phrases choisies de façon à ce que la cible soit précédée et suivie d'une suite parlée suffisamment étendue. Les durées totales et vitesses d'articulation (en syllabes par seconde) de divers fragments antérieurs et postérieurs aux cibles ont été mesurées. Les vitesses d'articulation ont été calculées après avoir retranché des durées totales les pauses non-articulatoires (supérieures à 100 ms). Les résultats (soumis ensuite à une analyse de variance) sont présentés figure 1.

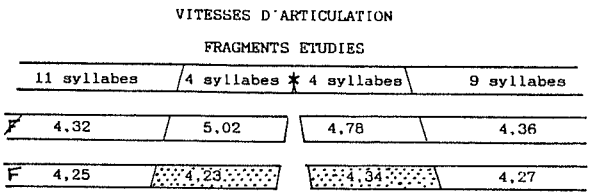


Figure 1: Ralentissement moyen des vitesses d'articulation.

Les ralentissements de la vitesse d'articulation étaient significatifs 4 syllabes avant et 4 syllabes après la cible pour chacun des 2 locuteurs. Les deux locuteurs indiquent donc la présence d'une cible environ 4 syllabes avant et après celle-ci, et de façon similaire. Une fois cette cible prononcée, c'est après 4 syllabes environ que le locuteur 1 retrouvait une énonciation similaire à celle qu'il avait en version monotone, alors que le locuteur 2 continuait à allonger sa durée d'énonciation, non pas en modifiant sa vitesse d'articulation, mais en réalisant davantage de pauses. Nous retrouvons la différence entre les deux locuteurs quant à la "concentration" des allongements dans l'entourage immédiat de la cible.

2. LES CORRELATIONS

Sur l'ensemble des textes, nous avons d'abord étudié séparément chaque indice temporel: nombre et durées des pauses, ou durées d'énonciation. On peut supposer que ces durées ne sont pas gérées par le locuteur indépendamment les unes des autres dans la chaîne parlée et que les ralentissements ont une certaine inertie. Il en découle que des corrélations devraient apparaître entre les allongements, s'ils sont gérés de façon cohérente par les locuteurs, et s'ils correspondent à une même fonction. Une absence de corrélation supposerait au contraire soit un manque de continuité dans la stratégie de lecture, soit des fonctions différenciées remplies par les divers allongements. Nous avons donc recherché si des corrélations se manifestaient. Les matrices du tableau 2 présentent les valeurs du coefficient de corrélation de Bravais Pearson r, calculées entre les indices temporels principaux, pauses avant (PAV), pauses après (PAP), durée de la syllabe avant cible (SYLL)) et de sa voyelle (VOY). Les versions en relief sont présentées entre parenthèses, et les versions monotones sans parenthèses.

Tableau 2 : Corrélations

	PAV	PAP	SYLL	VOY
Locuteur 1				
PAV		.20	.423	.258
PAP	(.181)		.092+	.136+
SYLL	(.373)	(.14)+		.654
VOY	(.40)	(.115)+	(.483)	
Locuteur 2				
PAV		.192	.468	.636
PAP	(.214)		.14+	.215
SYLL	(.458)	(.131)+		.777
VOY	(.539)	(.197)	(.669)	

Les coefficients r calculés témoignent d'une corrélation toujours significative entre les échelles de durées (prises deux à deux) des "syllabes", "voyelles" et "pauses avant", tous éléments qui précèdent la cible, aussi bien en version monotone qu'en version en relief pour les deux locuteurs. Sur les 24 coefficients calculés, seuls 6 montrent une absence de corrélation (codée + dans le tableau) : chaque fois c'est la liaison de l'un des éléments antérieurs à la cible avec la pause postérieure à la cible qui est concernée. La durée des pauses postérieures aux cibles n'est donc ni liée au ralentissement d'ensemble de la parole (car dans ce cas la corrélation serait positive et significative), ni utilisée pour compenser un ralentissement insuffisant (car dans ce cas, la corrélation serait négative et significative). Il n'est alors pas exclu que les pauses qui suivent la cible se distinguent des indices temporels précédents par leur fonction: alors que ce qui précède la cible témoignerait des processus développés par le locuteur pour se préparer lui-même à la focalisation, les pauses suivant la cible pourraient jouer le rôle de signal spécifiquement destiné à l'auditeur pour qu'il intègre le focus.

3. PROBLEMES SEMANTIQUES.

On sait que la réalisation intonative d'un focus dépend de son statut dans les hiérarchies syntaxique et énonciative (Rossi, 1985, 1987). Dans la lecture continue de textes, les réalisations temporelles en témoignent-elles ? Afin de répondre à cette question, deux textes T1 et T2 ont été sélectionnés. Un examen des cibles qu'ils comportaient conduisait intuitivement à distinguer deux catégories: des "concepts" (noms, verbes) importants pour le sens du texte, et des "modulateurs" de ces concepts (adverbes, adjectifs), plus secondaires pour la compréhension. Cette distinction permettait de classer ainsi les cibles: T1 : 3 concepts et 11 modulateurs, et T2 : 7 concepts et 7 modulateurs, (donc un plus grand nombre de concepts "importants" pour T2 que pour T1). Ceci constituait la phase A. Mais la simple nature grammaticale des mots (adjectifs, noms, adverbes ...) ne rend peut-être pas bien compte de la fonction que ces mots détiennent dans le texte. Pour trouver une typologie plus fondée psychologiquement, nous avons ensuite changé de critère : nous nous sommes basées sur le jugement d'auditeurs "naïfs" concernant l'importance relative des mots du texte. Les deux textes, lus sous forme monotone, ont été présentés à 92 auditeurs, qui avaient comme consigne de les prendre en note. Le dépouillement de ces notes a permis de faire un classement par rang de chacun des mots des textes en fonction de leurs fréquences relatives de rappel. Sur la base de ces classements, d'autres cibles ont été redéfinies en vue d'un ré-enregistrement des mêmes textes par le même locuteur: pour T1, les 14 cibles choisies occupent les premiers rangs de rappel (donc sont supposées être des concepts importants pour l'auditeur), et pour T2, les 14 cibles choisies occupent les derniers rangs (donc elles sont supposées secondaires pour l'auditeur). Ceci constitue la phase B.

Ainsi, deux types de critères ont présidé aux choix des cibles dans les phases A et B, et de plus, l'affectation des cibles "importantes" ou "secondaires" à chacun des deux textes est inversée entre ces phases. Le même locuteur (locuteur 1) a enregistré une version monotone et deux versions en relief, A et B, des deux mêmes textes. Nous avons ensuite mesuré les pauses précédant les cibles (ou, pour la version monotone, les pauses précédant les mots qui deviendraient des cibles en version en relief), les pauses suivant les cibles, les durées d'énonciation des cibles. Nous avons également voulu avoir un indice de la fréquence de la voix. Nous faisions l'hypothèse que le locuteur prononcerait de façon plus contrastée (certaines sons voisés plus graves et d'autres plus aigus) les cibles en version en relief. L'écart-type des fréquences fondamentales des sons voisés des cibles est alors notre quatrième variable dépendante. Pour chaque variable dépendante, nous avons comparé les versions en relief des textes, au sein de chaque phase pour deux textes différents et entre les deux phases pour un même texte. Aucune de ces comparaisons n'était significative, ni au sein des phases A ou B, alors que les cibles et les textes T1 et T2 se différenciaient par leur contenu sémantique, ni entre les phases A et B, alors que les deux versions en relief d'un même texte ne différaient que quant à "l'importance sémantique" relative des cibles sélectionnées. Tout se passe donc comme si les modifications prosodiques effectuées par le locuteur étaient toujours du même ordre, quelle que soit la signification intrinsèque des cibles et leur valeur sémantique au sein d'un texte.

Dans cette situation où l'ensemble des phrases du texte doit s'enchaîner continûment, il se peut que le statut du mot à focaliser soit sans effet temporel. Si les contraintes articulatoires et communicatives sont moins fortes (courtes phrases par exemple), il nous semble que la nature du focus doit se répercuter sur les paramètres prosodiques. La deuxième expérience reprend alors ce problème dans des conditions différentes.

EXPERIENCE 2.

Dans de courtes phrases (sujet, verbe, complément d'objet direct, adverbe, complément circonstanciel), les mots à focaliser pouvaient être soit l'adverbe (A) soit le complément d'objet (O), et leur ordre de succession variait (A puis O ou O puis A). Dans une condition "contrôle", les 8 locuteurs étudiés répètent, sans accent d'insistance, une phrase entendue juste avant. Deux conditions expérimentales sont constituées par 1) la lecture d'une phrase où le focus est écrit en majuscule; 2) la répétition d'une phrase entendue mais après qu'une question, comportant les termes interrogatifs "quoi" ou "comment", ait induit une focalisation sur le complément d'objet ou l'adverbe - technique utilisée par Pasdeloup (1984), Cooper & al. (1985), Eady & Cooper (1986).

En moyenne, quelque soit le focus, sa durée s'allonge de 25 % environ, le nombre et la durée des pauses le précédant et le suivant augmentent, enfin le pic de F0 s'élève de 45 Hz, mais le pic d'intensité ne change pas significativement. Le contour mélodique a été mesuré par un indice supra-segmental emprunté à Eady & Cooper (1986). La durée de l'intervalle entre le début du mot et la position du pic de F0 est rapportée à la durée totale du mot : une valeur comprise entre 0 et 0,5 indique un contour descendant, entre 0,5 et 1 un contour montant. La figure 2 présente schématiquement la succession des constituants : sujet (S), verbe (V), complément d'objet (O), adverbe (A) et complément circonstanciel (CC), et les contours mélodiques observés. En version sans focus (F), le contour dépend de la position du mot : il est descendant en première position et montant en deuxième. Par contre l'accentuation du focus (F) conduit à un contour toujours montant pour l'adverbe et toujours descendant pour le complément d'objet.

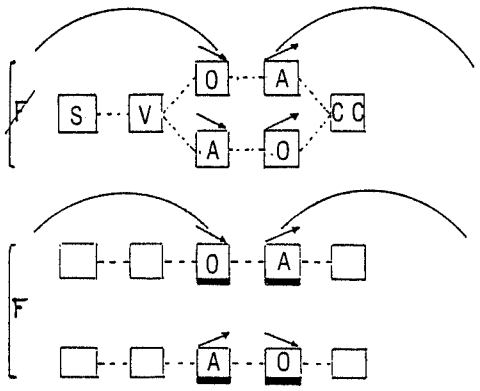


Figure 2: Contours mélodiques.

Concluons sur les processus cognitifs en cause dans les situations que nous avons étudiées. La focalisation d'un mot dans un énoncé implique une analyse syntaxique et sémantique de l'énoncé, une sélection et une identification du statut syntactico-sémantique du focus à transmettre, et une "promotion" du focus à destination de l'auditeur par divers moyens prosodiques. Dans la lecture continue de textes, certaines de ces étapes de traitement peuvent être insuffisamment effectuées et/ou marquées prosodiquement. Ainsi, les modifications temporelles importantes enregistrées au cours de la lecture continue dans l'expérience 1 ne sont pas modulées par le statut du focus. Si le locuteur doit effectuer par lui-même une sélection du focus, ou si la lecture s'effectue sur des phrases courtes, comme dans l'expérience 2, toutes ces étapes sont sans doute effectuées et marquées prosodiquement. Cependant, la marque prosodique n'est pas forcément pluri-paramétrique, elle peut porter préférentiellement sur un indice, comme le contour, et moins (ou pas du tout) sur un autre, comme le temps.

* Cette première expérience a été subventionnée par le CNET dans le cadre de la convention de recherche n° 89 1B226.

REFERENCES

COOPER, W.E., EADY, S.J.& MUELLER, P.R. (1985). Acoustical aspects of contrastive stress in question-answer contexts. *Journal of the Acoustical Society of America, 77*, 2142-2156.

CUTLER, A. (1987). Speaking for listening. In A. Allport, D.G. MacKay, W. Prinz & E. Echeerer (Eds), *Language Perception and Production. Relationships between listening, speaking, reading and writing* (pp 23-40). London, Academic Press L T D.

CUTLER, A. & ISARD, S.D. (1980). The production of prosody. In B. Butterworth (Ed), *Language Production. Volume 1 : Speech and Talk*, New-York, London, Academic Press.

EADY, S.J. & COOPER, W.E. (1986). Speech intonation and focus location in matched statements and questions. *Journal of the Acoustical Society of America, 80*, 402-415.

GERARD, C., DAHAN, D. & RIGAUT, C. (1991). *Etude des pauses : Relations entre la typographie d'un texte et sa structuration temporelle lors de la lecture.* Rapport de la convention de recherche n° 89 1B226 avec le CNET.

HUGGINS, A.W.F. (1978). Speech timing and intelligibility. In J. Requin (Ed), *Attention and Performance*, VII (pp 279-297). Hillsdale, Erlbaum Associates.

NOOTEBOOM, S.G., BROKX, J.P.L. & ROOIJ de, J.J. (1978). Contributions of prosody to speech perception. In W.J.M. Levelt & G.B. Flores d'Arcais (Eds), *Studies in the Perception of Language*. New York, Chichester, John Wiley.

PASDELOUP, V. (1984). *Etude acoustique et perceptive de la mise en valeur dans la phrase assertive en français*. Mémoire de Maîtrise de Linguistique, option phonétique, sous la direction de R. Gsell, Université Paris III.

ROSSI, M. (1985). L'intonation et l'organisation de l'énoncé. *Phonetica, 42*, 135-153.

ROSSI, M. (1987). Peut-on prédire l'organisation prosodique du langage spontané ? In Aspects prosodiques de la communication, *Etudes de Linguistique Appliquée, 66*, 20-48.

SEGUINOT, A. (1977). L'accent d'insistance en français standard. In F. Carton, D. Hirst, A. Marchal, A. Séguinot (Eds). *L'accent d'insistance. Emphatic stress* (pp 1-58). Didier, Studia Phonética (12).

SORIN, C. (1989). Perception de la Parole. In M.C. Botte, G. Canévet, L. Demany & C. Sorin (Eds), *Psychoacoustique et Perception Auditive* (pp 123-139). Paris, Inserm/SFA/CNET, Série Audition.

19es J.E.P. - Bruxelles, 19 au 22 mai 1992.

SCHWA, JONCTION ET DISJONCTION : SCHÈMES PROSODICO-SEGMENTAUX

Guéorgui JETCHEV

UNIVERSITÉ PARIS 7 - UNIVERSITÉ DE SOFIA

Résumé

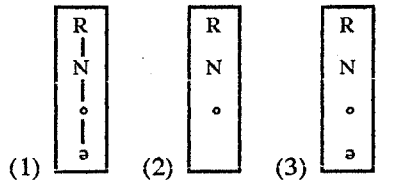
Cet exposé représente une interprétation phonologique des données de l'analyse statistique d'un corpus spontané présentées dans Jetchev 1991a [1], fondée sur la modélisation élaborée dans Jetchev 1991b [2]. Les grandes lignes du modèle et les principes de catégorisation des schwas font l'objet des paragraphes 1.0-1.3. La suite est consacrée à la caractérisation des deux contextes où apparaissent des schwas intermittents autour des frontières de syntagmes prosodiques. Ces schwas témoignent d'une tendance nette à la manifestation dans certains contextes et à la non manifestation dans d'autres contextes prosodico-segmentaux. Les schèmes de réalisation des schwas sont intérprétés en tant que véhicules de la fonction configurative et marques de jonction/disjonction. Il s'avère que le schème jonctif va de pair avec le non-isomorphisme des niveaux lexical et post-lexical, alors que le schème disjonctif implique des structures isomorphes à ces deux niveaux.

1. En français non méridional, les schwas de certaines catégories constituent des points de variabilité dans les représentations lexicales. Dans le cadre d'un modèle plurilinéaire (du type du modèle tridimesionnel de P.Encrevé [3]), ces schwas correspondent à des sites phonologiques spécifiques : segments flottants ou vides sur la ligne autosegmentale des segments (positions flottantes selon J.Goldsmith [4]). Dans ces sites on constate une fluctuation des usages phonétiques, en l'occurrence des réalisations des schwas (manifestation ou non manifestation), qui n'assument jamais de fonction distinctive et, par conséquent, se prêtent souvent à l'expression de traits redondants, configuratifs, stylistiques ou situationnels (cf. [5], [6]).

Le modèle présenté dans [2] vise à rendre compte des réalisations concrètes des schwas en discours spontané.

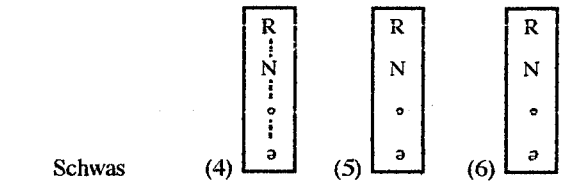
L'analyse des représentations phonologiques s'y fait sur deux **niveaux** (lexical et post-lexical), à la suite desquels les schwas appartenant à certaines catégories traversent un ou plusieurs **filtres** (rythmique ou métrique, sociosituationnel, débit de l'énonciation, pragmatique).

1.1. A la sortie du niveau lexical les schwas appartiennent à une des trois catégories suivantes : schwas **ancrés** dans le squelette (*entr<u>e</u>tien, fermet<u>é</u>, br<u>e</u>bis)* ; schwas **épenthétiques** (*souv<u>e</u>rain; grand<u>e</u>, ami<u>e</u>, pancart<u>e</u>*) ; schwas **flottants** (*n<u>e</u>veu ; l<u>e</u>*), auxquelles corespondent les représentations lexicales (1), (2), (3).



schwas ancrés — schwas épenthétiques — schwas flottants

1.2. Si les schwas ancrés ne subissent pas de modifications au niveau post-lexical, le sort des schwas flottants et d'une partie des schwas épenthétiques (en finale de polysyllabes) se décide essentiellement à ce niveau-là. Une partie des schwas flottants sortent du niveau post-lexical comme ancrés dans le squelette (ceci est régi par les rapports de gouvernement entre finales et initiales de mot qui s'établissent au niveau post-lexical), mais d'autres le traversent tout en restant flottants, c-à-d. potentiellement sujets à variation. De même, une partie des schwas épenthétiques restent tels à l'issue du niveau post-lexical, tandis que d'autres deviennent flottants. Nous adopterons le terme de **schwas intermittents** pour désigner tous les schwas qui sortent en tant que flottants du niveau post-lexical. Ce sont les schwas qui correspondent aux représentations post-lexicales (5) et (6).



(4) *quel neveu, il aime le café, le héros, pas de hache*

(5) *mon neveu, devant chez moi, prends le café, de mon côté*

(6) *pancarte bleue, quelle hauteur*

La distinction des catégories (1), (2) et (3) découle des rapports de gouvernement inter-segmental au niveau lexical avec la prise en compte des frontières de mots (monosyllabes; polysyllabes : syllabe interne, initiale, finale) et du degré de "vitalité" des schwas, en ce qui concerne la distinction entre (2) et (3).

La distinction des trois sous-catégories de schwas flottants (4), (5) et (6) découle des rapports de gouvernement inter-segmental qui s'établissent aux frontières de mots au niveau post-lexical.

1.3. Le modèle ainsi dégagé présuppose que l'analyse des réalisations des schwas en discours spontané se fasse en trois temps : (a) au niveau des représentations lexicales, (b) au niveau post-lexical, (c) en surface. Les différences constatées entre les représentations des schwas en (b) et en (c) sont dues à l'influence des filtres. A titre d'illustration, nous proposons l'analyse d'une séquence de notre corpus (fig. 1).

Dans l'exemple de la fig.1, la seule différence entre (b) et (c) est en e_6. Elle doit être interprétée comme résultant de l'influence d'un filtre (en l'occurrence, du filtre débit de l'énonciation).

2. Les schèmes jonctif et disjonctif de réalisation des schwas constituent dans le cadre de ce modèle l'un des filtres, que l'on pourrait appeler "lien syntagmatique jonctif/disjonctif"[1]. Ce filtre ne semble avoir de l'incidence que sur les schwas intermittents, à la différence, par exemple, du filtre pragmatique qui affecte aussi les schwas catégorisés comme épenthétiques.

On a pu établir les schèmes privilégiés de réalisation (manifestation/non manifestation) des schwas en fonction du contour rythmico-mélodique de l'énoncé, qui peut marquer un rapport syntagmatique de jonction ou de disjonction entre les constituants prosodiques successifs. Ces schèmes peuvent donc être considérés en tant que véhicules de la fonction configurative, de la démarcation.

Afin de dégager le rôle du filtre "lien syntagmatique jonctif/disjonctif", les réalisations des schwas intermittents dans deux contextes ont été prises en compte :

2.1. Contexte 1 (C1): Schwas intermittents en syllabe initiale de syntagme prosodique, le syntagme prosodique précédent étant à finale vocalique.

L'étude spectrographique de Rialland [7], portant entre autres sur des schwas de ce type, qu'elle désigne comme des *schwas nucleus*, révèle qu'un certain nombre d'indices acoustiques nous permettent de considérer que la non manifestation de ces schwas (cf. [7 : 201-202] : *on va t(e) renverser*, avec chute du schwa [a-t–rã], comparé à *on va traverser* [a-tra]) n'aboutit pas à une resyllabation de la consonne vers la droite. Qui plus est, il y a lieu d'admettre que lorsque la suite Cə se trouve en **C1** (cf. un exemple analogique à celui de Rialland, où le schwa de *te* est en **C1** : *il voulait absoluMENT | t(e) renverSER; il voulait absoluMENT | traverSER)*, la syllabe accentuée qui précède immédiatement, marquée ici en majuscules, exerce une force d'attraction sur la C en cas de chute du schwa et tend à en faire sinon une coda, au moins une consonne ambisyllabique : [ˊã-t-rã][2]. Nous serions donc là en présence d'une resyllabation à gauche. La consonne (entièrement ou partiellement) resyllabée à gauche est à la base du **non-isomorphisme** que l'on peut constater ici entre l'analyse au niveau **lexical** et l'analyse au niveau **post-lexical**. Sur le plan de la structure lexico-grammaticale, le [t] de *t(e)* appartient au syntagme morpho-syntaxique qui suit (M2) alors que sur le plan de la structure syllabique et prosodique, par suite de la resyllabation qui se produit au niveau post-lexical, ce [t] s'intègre aussi au syntagme prosodique qui précède (P1), cf. fig.2.

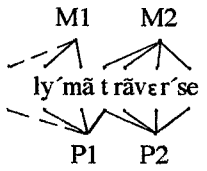


fig.2

[1] Ce terme nous semble préférable à celui utilisé dans [2] : *organisation hiérarchique de l'énoncé.*

[2] Pour ce qui est de l'ambisyllabicité d'une consonne à l'intervocalique et son rapport avec la place de l'accent , cf. **Anderson**, J.M. & **Ewen**, C.J.,1987 : *Principles of Dependency Phonology*, Cambridge, University Press : 66.

Le non-isomorphisme des deux niveaux, dû à l'enchaînement post-lexical, nous semble fonctionner comme une marque de jonction pour les locuteurs-auditeurs. Au contraire, le maintien du schwa en **C1** rendrait impossible un tel effet d'enchaînement et évoquerait, par conséquent, un lien syntagmatique disjonctif.

2.2. Contexte 2 (C2): Schwas intermittents en syllabe finale de groupe prosodique, post-tonique[3], le groupe prosodique suivant étant à initiale consonantique.

On va se servir du terme *appendice féminin*, proposé par M.Plénat [8] pour désigner ce type de "syllabe" (cf. aussi le terme *mot phonétiquement féminin*, introduit par F.Dell [9], et qui désigne un mot se terminant par un appendice féminin). Il y a des indices en faveur de l'appartenance de l'appendice féminin à la structure précédente et d'autres qui révèlent plutôt son appartenance à la structure subséquente.

D'après les paramètres accentuels du français, tels qu'ils ont été définis par D. Hirst [10 : 152][4], l'appendice féminin en **C2** se trouve dans un pied prosodique, tête à gauche, avec la syllabe précédente.

Du point de vue de la formation des hypocoristiques à redoublement (cf.[8]), l'appendice féminin semble former un tout indivisible avec la consonne codique précédente : ou bien il disparaissent ensemble *(Georges –> jojo)*, ou bien ils sont maintenus ensemble *(Adolphe –> dodolphe)*. Plénat [8 : 172-173] les analyse donc dans le cadre de la rime de la syllabe précédente (une sorte d'"hypersyllabe").

P. Mertens [11 : 84] signale que lorsque la syllabe accentuée d'un mot phonétiquement féminin est porteuse d'un ton dynamique, la réalisation de la deuxième more de ce ton est souvent partiellement ou même entièrement répartie sur l'appendice féminin. La réalisation phonétqiue de ce dernier implique le maintien du schwa intermittent. Du point de vue de la courbe mélodique, l'appendice féminin ferait donc partie de l'unité intonative (UI) précédente.

D'autre part, selon Cornulier [12 : 112], un appendice féminin s'intègre rythmiquement à l'unité de sens qui le suit, même parfois s'il en est séparé par une pause. Il semble être pris en compte par les auditeurs français dans le rythme du syntagme prosodique (SP) subséquent, cf. fig.3.[5]

[3] On va se servir par la suite du terme *appendice féminin*, proposé par M.Plénat 1984 pour désigner ce type de "syllabe". Cf. aussi le terme *mot phonétiquement féminin*, introduit par F.Dell 1984 et qui désigne un mot se terminant par un appendice féminin.

[4] (1) extramétricité : aucune ; (2) gabarit du pied prosodique : [S (C_o ə)] ; (3) dominance : initiale dans le pied : [´ _] , finale dans le mot : { [ ] [ ] ... [´]}

Les paramètres accentuels de Hirst nous semblent caractériser le niveau lexical : à ce niveau-là *-tre* sera analysé dans une même unité (pied prosodique) avec la syllabe précédente *qua-*. Pour ce qui est du niveau post-lexical, *-tre* y sera inclus dans le cadre du groupe accentuel[7] (GA) qui le suit *(-tre les GROS* ou *-tre les gros coCHONS*, suivant que *gros* reçoit ou non un accent rythmique primaire ou secondaire)[7].

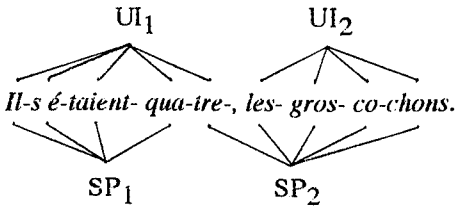


fig. 3

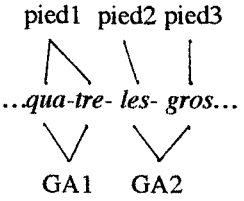


fig.4

La syllabe à [ə] posttonique apparaît donc comme faisant partie du même pied prosodique que la syllabe précédente au niveau lexical (c'est probablement la raison pour laquelle un ton dynamique se répartit sur ces deux syllabes), mais aussi comme partie intégrante du patron rythmique du syntagme prosodique suivant.

Le **non-isomorphisme** constaté entre le niveau lexical et le niveau post-lexical quant à l'appartenance de l'appendice féminin C_1ə (analysé dans l'unité à gauche au niveau lexical, dans l'unité à droite au niveau post-lexical) nous semble pouvoir expliquer le fonctionnement de C_1ə comme un indice de **jonction** par analogie avec la consonne ambisyllabique en **C1**. Pour qu'il y ait non-isomorphisme, et donc jonction, le schwa intermittent de C_1ə doit se manifester en surface. Si le schwa de C_1ə tombe (si on prononce *quat-* au lieu de *quatre* dans l'exemple ci-dessus), la marque jonctive ne pourra pas être réalisée. De même, si en **C1** le schwa est maintenu, il n'y aura pas d'ambisyllabicité

[5] Pour ce terme et son application en prosodie française voir [13].

de la consonne précédente et, par conséquent, les niveaux lexical et post-lexical seront parfaitement isomorphes.

2.3. L'interaction constatée entre les réalisations des schwas dans les contextes **C1**, **C2** et les structures prosodiques qui les accompagnent, nous permet de définir l'existence en français de deux schèmes prosodico-segmentaux (cf. fig.5).

Le **schème jonctif** implique une tendance à la non manifestation des schwas intermittents en **C1** et à leur manifestation (**schwas jonctifs**) en **C2**. Dans les deux cas, la jonction est supposée être signalée par le non-isomorphisme entre le niveau lexical et le niveau post-lexical (consonne ambisyllabique dans le cas du **C1** et syllabe ambiconstituante, analysée tantôt à gauche, tantôt à droite dans le cas du **C2**).

En revanche, le **schème disjonctif**, qui implique une rupture nette dans les valeurs des paramètres prosodiques, tend à favoriser l'isomorphisme des deux niveaux. Par conséquent, avec ce schème, les schwas intermittents en **C1** ont tendance à se manifester (**schwas disjonctifs**), permettant ainsi à éviter une consonne ambisyllabique, alors que ceux du **C2** témoignent d'une nette tendance à la non manifestation, ce qui exclut la réalisation d'un appendice féminin, structure impliquant le non-isomorphisme.

[1] **Jetchev**, G. , 1991a : "Marques segmentales sociosituationnelles en français contemporain", *Actes du XIIème Congrès International des Sciences Phonétiques*, 3, pp.146-150.

[2] **Jetchev**, G. , 1991b : *Essai de modélisation des catégories de schwas en français non méridional. Application à l'étude de corpus spontanés*, Université Paris VII, DRL, mémoire de DEA.

[3] **Encrevé**, P., 1988 : *La liaison avec et sans enchaînement. Phonologie tridimensionnelle et usages du français*, Paris, Seuil.

[4] **Goldsmith**, J. A., 1990 : *Autosegmental and Metrical Phonology*, Oxford, Basil Blackwell.

[5] **Léon**, P.R., 1987 : "E caduc : facteurs distributionnels et prosodiques dans deux types de discours", *Proceedings XIth ICPhS*, v. 3, pp. 109-112.

[6] **Lucci**, V., 1983b : *Etude Phonétique du Français Contemporain à travers la Variation Situationnelle (débit, rythme, accent, intonation, e muet, liaisons, phonèmes)*, Publications de l'Université des Langues et Lettres de Grenoble.

[7] **Rialland**, A., 1986 : "Schwa et Syllabes en Français", *in* **Wetzels**, L. & **Sezer**, E., eds., *Studies in Compensatory Lengthening*, Dordrecht, Foris, pp. 187-226.

[8] **Plénat**, M., 1984 : "Toto, Fanfa, Totor et même Guiguitte sont des ***anars***", in **Dell**, F., **Hirst**, D. & **Vergnaud**, J.-R., eds., *Forme sonore du langage*, Paris, Hermann, pp.161-181.

[9] **Dell**, F., 1984 : "L'accentuation dans les phrases en français", *in* **Dell**, F., **Hirst**, D. & **Vergnaud**, J.-R., eds., pp.65-122.

[10] **Hirst**, D., 1987 : *La représentation linguistique des systèmes prosodiques: une approche cognitive*, thèse d'Etat, Aix-en-Provence, Université de Provence.

[11] **Mertens**, P., 1987 : *L'intonation du français. De la description linguistique à la reconnaissance automatique*, doctorale dissertatie, Katholieke Universiteit Leuven.

[12] **Cornulier**, B. de, 1982 : *Théorie du vers. Rimbaud, Verlaine, Mallarmé*, Paris, Seuil.

[13] **Di Cristo**, A., in press : "Intonation in French", *in* **Hirst**, D. & **Di Cristo**, A., eds., *Intonation Systems : A Survey of Twenty Languages*, Cambridge, Cambridge University Press.

j 'a´vais:# a´ssez de$_1$ bon ´´sens pour m'aperce$_2$´´voir que$_3$ `ça ne$_4$ me$_5$ me$_6$´nait ` pas très ´´loin # [1]

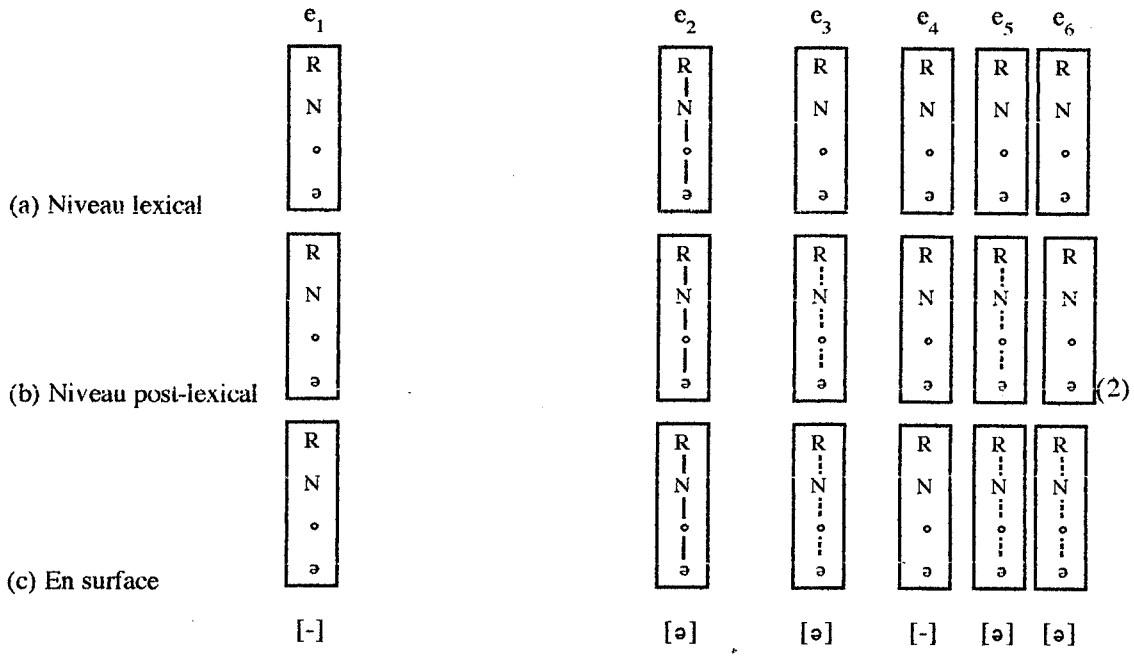


fig.1

	Schème ***jonctif***	Schème ***disjonctif***
Contexte 1	Tendance à la non manifestation du schwa ↔ *consonne ambisyllabique*	Tendance à la manifestation du schwa ⇓ **schwas disjonctifs**
Contexte 2	Tendance à la manifestation du schwa ⇓ **schwas jonctifs** ↔ *syllabe ambiconstituante*	Tendance à la non manifestation du schwa
	Non-isomorphisme	**Isomorphisme**

(des niveaux lexical et post-lexical)

fig.5

[1] Claude Lévi-Strauss, Radioscopie, Cassettes Radio France, K 1243, 1989; 1-189.

(2) Du moment que le schwa de "me" a été maintenu, celui de "menait" est considéré comme un schwa intermittent noté "**e**"; sinon il aurait été noté "**e̲**" comme un schwa ancré post-lexicalement. Tel est d'ailleurs le cas du schwa maintenu de "me" qui, à l'origine intermittent comme tout schwa de monosyllabe, par suite de la chute du schwa dans la syllabe précédente "ne", s'ancre dans le squelette au niveau post-lexical.

19es J.E.P. - Bruxelles, 19 au 22 mai 1992.

VERS UNE REPRESENTATION AUTOSEGMENTALE DE L'ACCENT ITALIEN: ETUDE EXPERIMENTALE

CARMELA SPAGNOLETTI

FNRS — UNIVERSITE LIBRE DE BRUXELLES

Résumé

Dans cet article sont présentés les résultats d'une expérience visant à évaluer la nature autosegmentale de l'accent italien. Afin de déterminer si l'accent est une propriété du mot, ou au contraire, d'une syllabe particulière, j'ai enseigné à des locuteurs natifs de l'italien un jeu de langage, ou langue secrète, qui consiste à bouleverser l'ordre des syllabes d'un mot. Le test comprend des mots bisyllabiques paroxytons, ainsi que des trisyllabes paroxytons et proparoxytons. Si l'accent est une propriété intrinsèque d'une syllabe particulière, les caractéristiques accentuelles de cette syllabe se déplaceront automatiquement avec le reste de ses composantes phonétiques lors du changement d'ordre des syllabes. Or les résultats obtenus montrent clairement qu'il est possible d'agir **indépendamment** sur les syllabes et sur le patron accentuel. L'accent peut dès lors être considéré comme une unité autonome par rapport aux autres composantes phonologiques du mot.

INTRODUCTION

Dans le but de déterminer si l'accent est une propriété du mot ou, au contraire, d'une syllabe particulière, j'ai enseigné un jeu de langage à des locuteurs natifs de l'italien[1]. Il eût été préférable de trouver des locuteurs qui pratiquent habituellement un jeu de langage intéressant pour notre objet d'investigation. Malheureusement, je n'en ai rencontré aucun. Cependant, la méthode qui consiste à apprendre à des locuteurs à faire une manipulation sur des éléments de leur langue correspond à ce que l'on appelle communément une "expérience". Et si la plupart des sciences ont depuis bien longtemps adopté la méthode expérimentale, tant celle-ci se révèle efficace pour l'évaluation des hypothèses théoriques, la phonologie, quant à elle, n'en est aujourd'hui encore qu'à ses débuts. Mais ces débuts sont prometteurs si l'on en croit la multiplication des recherches dans cette voie (cf. notemment l'ouvrage de J. Ohala & J. Jaeger (1986)). Mon propos n'étant pas de faire un plaidoyer pour la phonologie expérimentale, je me contenterai de renvoyer le lecteur à l'ouvrage mentionné ci-dessus, ainsi qu'à J. Ohala (1984), (1986), et P. M. Bertinetto (1988).

ÉTUDE EXPÉRIMENTALE

Dans le but de mettre à l'épreuve l'hypothèse selon laquelle l'accent est une propriété du mot et non d'une syllabe particulière, j'ai enseigné aux locuteurs testés quelques aspects d'un jeu de langage, ou langue secrète, qui se pratique en Afrique (cf. Demolin 1991) et qui consiste à bouleverser l'ordre des syllabes. Ainsi, en ce qui concerne les mots bisyllabiques, la séquence S1 S2 devient S2 S1. En ce qui concerne les polysyllabes, plusieurs modifications sont possibles. Parmi celles-ci, j'ai retenu l'inversion des trisyllabes S1 S2 S3 en S3 S1 S2. Ce jeu de langage (spontané chez ses locuteurs) a permis à D. Demolin de déterminer, entre autres choses, que le schème tonal est une propriété du mot en mangbetu, car il reste inchangé bien que les syllabes aient été permutées. J-M. Hombert (1986) a également enseigné, à des locuteurs de diverses langues à ton, les deux permutations suivantes: C1V1C2V2 > C1V2C2V1 et C1V1C2V2 > C2V2C1V1. Les résultats obtenus suggèrent que parmi les diverses langues à tons ainsi étudiées, le schème tonal est une propriété du mot dans les langues africaines (bakwiri, dschang, kru), mais pas dans les langues asiatiques (mandarin, cantonais, taiwanais et thai) où les réponses des locuteurs sont plus variées, comprenant des cas où le ton se déplace avec la voyelle ou la syllabe. Hombert suggère que cette différence de traitement des tons pourait être due, en partie, au fait que ces deux groupes de langues sont de types

[1]J'ai été aidée dans ma tâche par Sita Trini Castelli, qui s'est aimablement proposée pour recueillir des données, et qui a notemment testé des locuteurs lors d'un voyage en Italie.

morphologiques très différents. Afin d'obtenir des réponses spontanées quant au placement des tons, Hombert à donné à ses locuteurs des exemples de mots où le ton est identique sur les deux syllabes à manipuler. Viennent ensuite les items expérimentaux dans lesquels chaque syllabe porte un ton différent. Dans le cadre d'une étude comme celle-ci, qui porte sur l'accent, se pose évidemment le problème de l'entraînement. En effet, tout mot possède au moins un et un seul accent principal. Il est donc impossible de prononcer des exemples sans influencer d'une façon ou d'une autre la réponse du locuteur. Par contre, l'orthographe italienne ne note pas la place de l'accent sauf dans le cas où celui-ci frappe la syllabe finale (patron oxyton). Il est dès lors possible, voire même nécessaire, qu'une partie de l'expérience se fasse par écrit.

SUJETS

Les sujets sont des locuteurs natifs de l'italien qui se sont prêtés volontairement à ce petit jeu. Ce groupe se compose de 7 femmes et 5 hommes, âgés de 19 à 45 ans et un sujet de 84 ans. Ces 12 sujets sont originaires du nord et du centre de l'Italie et parlent tous une variante standard de l'italien, sans accent régional prononcé. Les variations de prononciation inévitables entre les locuteurs de ces deux régions ne sont pas pertinentes pour la nature de l'étude en question. Le corpus initial comprenait deux sujets supplémentaires, qui ont été écartés de l'analyse pour diverses raisons. Remarquons enfin, que 5 sujets ont été testés lors de leur séjour en Belgique (étudiants du cycle supérieur universitaire (2e et 3e cycle) et non universitaire), les 7 restants ont été interrogés en Italie (à Milan) (ce sont principalement des employés de bureau de divers niveaux, ainsi que des personnes exerçant des professions libérales). Une partie des données fut récoltée par moi-même, l'autre par une aimable collaboratrice qui a sacrifié à notre cause une partie de ses vacances en Italie.

MÉTHODE

Les sujets sont informés qu'on va leur apprendre un jeu de langage, ou langue secrète, qui consiste à manipuler des mots d'une façon précise. Ils reçoivent un formulaire qui se compose de trois "exercices". Dans l'exercice 1, on demande d'inverser les syllabes de façon à ce que l'ordre 1-2 devienne 2-1. L'instruction est suivie de 5 exemples de mots bisyllabiques paroxytons accompagnés de la réponse "correcte". Par exemple il cane > il neca. Suit une série de 10 mots tests bisyllabiques, avec un "blanc" pour écrire la réponse. Par exemple, il tipo > il ..... Les mots sont tous des substantifs de type morphologique simple (c'est-à-dire sans affixes dérivationnels) afin d'éviter une influence possible de la morphologie[2]. Ces mots, ainsi que les réponses, sont toujours précédés de l'article afin que le sujet garde à l'esprit qu'il utilise des mots et leur variante disons "déguisée", et non pas par exemple des séquences de syllabes sans signification. L'exercice 2 est organisé de façon similaire excepté que les mots comptent 3 syllabes et que le patron accentuel est paroxyton dans 8 cas, proparoxyton dans les 7 autres, ces deux patrons alternant dans un ordre aléatoire. L'instruction est cette fois de changer l'ordre des syllabes de façon à ce que l'ordre 1-2-3 devienne 3-1-2. Remarquons que tous les mots sont composés de syllabes de type CV, ceci afin de permettre des permutations aisées, sans risquer de créer des séquences qui violeraient la phonotactique de la langue. Lorsque le formulaire est complété, ce qui est fait très rapidement, le sujet lit à voix haute au total 6 fois tous les couples mot-test / mot-réponse, y compris les exemples. La raison de ces lectures multiples est de vérifier la cohérence entre les réponses que le sujet donne pour un même mot. Il est souvent arrivé que, s'interrompant subitement, le sujet pose explicitement la question de la place de l'accent. Peut-être est-ce là un indice que l'accent n'est pas lié intrinsèquement à une syllabe; sinon on comprend mal comment un choix pourrait être possible. L'ordre de lecture a été varié de façon à détecter une influence possible du patron d'un mot précédent sur celui du mot suivant. Ainsi, 1 sujet a effectué les 6 lectures du début à la fin du test; 4 sujets (soit le groupe A) ont lu 3 fois du début à la fin suivi de 3 fois de la fin au début; enfin, 7 sujets (soit le groupe B) ont lu alternativement du début à la fin du test et de la fin au début. Les lectures ont été enregistrées afin de permettre des transcriptions aisées et de pouvoir réécouter plusieurs fois le même mot, si besoin était. Les jugements sur la place de l'accent ont été faits à l'oreille par moi-même. Le test comporte en réalité un troisième exercice, composé de syntagmes nominaux complexes de divers types, ainsi que d'une phrase. Je n'ai cependant pas traité les données concernant cette troisième partie, car viennent se greffer sur l'accent lexical des phénomènes liés à l'accentuation de groupe, voire de phrase, et des phénomènes intonatoires, ce qui devient impossible à traiter à l'oreille de façon rigoureuse.

RÉSULTATS ET DISCUSSION

Nous disposons au total de 2160 réponses produites, soit 1080 mots bisyllabiques et autant de trisyllabes. En ce qui concerne les mots de 2 syllabes, tous les sujets, sans exception et sans variation de réponse d'une lecture à l'autre, ont conservé le patron paroxyton pour tous les mots présentés. En revanche, les mots de 3 syllabes sont traités d'une façon moins homogène, comme on peut le voir au tableau 1. Appelons M le type de réponse où le patron accentuel du mot-réponse est identique à celui du mot-test, et D,

[2]Rappelons en effet que l'italien étant une langue dite à accent libre, les morphèmes ont des propriétés accentuelles (cf. Garde 1968).

le type de réponse où les patrons sont différents. Ecartons d'emblée le problème de l'influence possible du mot qui précède sur celui qui suit. Lorsque l'ordre de lecture est inversé, le patron accentuel du mot qui précède peut être différent. Afin de tester l'hypothèse d'une influence éventuelle de l'ordre de lecture sur le choix de la réponse, j'ai comparé le nombre de réponses de type M dans les lectures début-fin et fin-début, dans le groupe A et le groupe B séparément. Pour le groupe A, un seul sujet sur les 4 a donné une réponse M supplémentaire dans l'ordre fin-début. Cette réponse différente ne peut cependant pas être imputée à l'influence du mot qui précède car dans ce cas précis, le mot précédent présente le même patron accentuel quel que soit l'ordre de lecture. En ce qui concerne le groupe B, sur les 7 sujets, 1 a donné le même nombre de réponse dans les deux types de lecture, 5 ont donné une réponse supplémentaire dans l'ordre fin-début, et 1 sujet a donné 8 réponses supplémentaires dans l'ordre début-fin. Encore une fois, l'observation des données brutes montre que le patron des mots contigus n'intervient en rien dans ces variations. Mis à part un cas atypique, la très légère variation dans le nombre de réponses M semble plutôt être due à l'effet du hasard, ou peut-être à la distraction et non pas commandée par un effet d'ordre. J'ai également comparé ces deux groupes, toujours sur base des réponses M. La différence entre le groupe A et le groupe B n'est pas statistiquement significative ("Two-tailed Mann-Whitney U test": U = 11,5 avec $n_1 = 4$ et $n_2 = 7$; $0.64 \leq p \leq 0.78$). En d'autres termes, cela signifie que d'une part, la variation dans le type de réponses que l'on observe n'est pas due à une influence mutuelle du patron accentuel des mots voisins, ni à la façon dont sont ordonnés les types d'ordre de lecture. Pour la suite de l'analyse, on peut dès lors fusionner les deux groupes et ajouter le sujet qui avait effectué les 6 lectures dans le même ordre. Ajoutons également qu'il n'y a pas de différence significative d'une lecture à l'autre. La comparaison a de nouveau été faite à partir du nombre réponses de type M, et pour les 12 sujets ("Friedman two-way analysis of variance by ranks": $\chi^2_r = 2.059$, df=5, $0.80 \leq p \leq 0.90$). Ces résultats confirment l'impression d'une certaine cohérence globale des réponses lorsque l'on regarde les données brutes. Et cependant, aucun sujet n'a fourni 100% de réponses de type M, contrairement à ce qui se passe pour les mots bisyllabiques. De plus, il n'y a pas de différence statistiquement significative entre le nombre de réponses M et le nombre de réponses D ("Wilcoxon matched-pairs signed ranks test": T = 20.5 avec N = 12). Si l'on compare le nombre de réponses M et le nombre de réponses où l'accent semble s'être déplacé avec la syllabe (soit le type mD), on n'obtient pas non plus de différence significative (T = 32, avec N = 11). La seule comparaison qui atteigne le seuil de signification est la différence entre le nombre de réponses mD et le nombre de réponses où le patron accentuel est différent de celui du mot-test mais où la syllabe accentuée est différente également (soit le type dD), avec une prépondérence des réponses mD (T = 6, avec N = 12, $p \leq 0.01$). A ce stade de l'analyse, le comportement des sujets n'est pas aisé à interpréter. En effet, on ne remarque aucune préférence entre les deux traitements: déplacer l'accent avec la syllabe qui le porte ou conserver le même patron accentuel sur le mot-réponse. On peut s'interroger sur les raisons de cette divergence de traitement selon que les mots sont de deux ou de trois syllabes. Or dans le corpus, tous les mots de deux syllabes sont paroxytons, et cela pour deux raisons. D'une part, l'orthographe italienne note l'accent final, ce qui aurait sans doute influencé les réponses (rappelons que le test comporte une partie écrite). D'autre part, il existe comparativement très peu de mots bisyllabiques oxytons, et — ce qui limite de façon dramatique les possibilités pour ce test — le nombre de substantifs dont la structure syllabique est CVCV est encore plus restreint. Tous les mots-test bisyllabiques possèdent donc le même patron accentuel. Pour les mots de trois syllabes, par contre, on dispose de nombreux mots proparoxytons et paroxytons répondant aux exigences du test. Il est intéressant de voir si ces deux patrons suscitent des réponses de types différents. L'analyse statistique effectuée sur le nombre de réponses de M paroxytons et M proparoxytons montre que c'est bien le cas (T = 6, avec N = 12; $p \leq 0.01$) et la différence va dans le sens M paroxytons > M proparoxytons. Si l'on regarde les données brutes, la tendence est nettement visible (cf. Tableau 1). En effet, 8 sujets sur les 12 n'ont donné **aucune** réponse M ayant le patron proparoxyton. Un fait encore plus intéressant: pour l'ensemble des 12 sujets, **toutes** les réponses mD ont le patron paroxyton, et par conséquent aucune réponse mD n'est proparoxyton. De même, aucune réponse dD n'a le patron paroxyton. En d'autres termes, lorsque le mot-test est paroxyton, ce patron est conservé dans la majorité des cas (71%), dans la minorité des cas, c'est le patron proparoxyton qui est "collé" au mot-réponse, et en aucun cas l'accent ne se déplace avec la syllabe qui était accentuée dans le mot-test, ce qui aurait donné un patron oxyton (S1'S2S3 > S3S1'S2). En revanche, lorsque le mot-test est proparoxyton, ce patron n'est conservé dans le mot-réponse que dans 18% des cas. De plus, ces 18% sont le fait de 4 sujets seulement sur les 12. Dans la grande majorité des réponses (82%), le patron du mot-réponse est paroxyton, et cela correspond également à l'interprétation selon laquelle l'accent serait déplacé avec la syllabe accentuée ('S1S2S3 > S3'S1S2). On observe donc une nette préférence pour le patron paroxyton: quel que soit le patron accentuel du mot-test, on obtient 77% de mots-réponses paroxytons (la proportion des mots-tests paroxytons est de 53%).

CONCLUSIONS

Les résultats d'une analyse détaillée des types de réponses est davantage compatible avec l'hypothèse selon laquelle le patron accentuel est une propriété du mot. Si l'accent avait été intrinsèquement lié à la syllabe accentuée, il se serait déplacé automatiquement avec elle. Or ce n'est jamais le cas avec les bisyllabes;

quant aux trisyllabes, l'accent se déplace apparemment avec la syllabe accentuée uniquement si cela donne un mot-réponse paroxyton. Lorsque cette démarche devrait donner un mot-réponse oxyton (cas des mots-tests paroxytons), cette possibilité n'est jamais choisie. Les sujets optent alors pour le maintien du même patron ou encore ils accentuent une syllabe qui n'était pas accentuée dans le mot-test et qui donne un patron paroxyton. Ce type de réponse serait difficilement compréhensible si l'accent était intrinsèquement lié à la syllabe. Si par contre le patron accentuel est une propriété du mot, ce type de réponse ne sort pas de la démarche générale des sujets, celle de choisir un patron accentuel pour le nouveau mot créé, indépendamment de la question de savoir quelle sera la syllabe qui portera les marques phonétiques de l'accent. Les sujets qui ont demandé comment ils devaient accentuer les mots-réponses, et à qui on a répondu de répondre ce qui leur vient spontanément à l'esprit, n'ont montré, par la suite, aucune hésitation. Une fois un patron choisi, celui-ci était généralement maintenu dans les lectures successives. Quand le test fut terminé, certains sujets ont commenté leur choix en disant qu'ils avaient parfois changé de patron accentuel, mais ne savaient pas pourquoi, d'autres ont déclaré s'être inspirés de la prononciation d'un mot existant qui ressemblait au mot-réponse. Il semble donc clair que pour les sujets, l'inversion des syllabes est une chose indépendante de la question de la place de l'accent.

Il reste deux problèmes à élucider. On peut en effet s'interroger sur le phénomène d'évitement apparent du patron oxyton. Rappelons que ce patron est minoritaire en italien, et qu'en général il est le résultat des propriétés accentuelles de certains morphèmes. Pour les substantifs, il s'agit d'ailleurs de mots qui exceptionnellement ne possèdent aucune flexion, d'où leur dénomination, en grammaire traditionnelle, de mots "tronqués". De plus, comme je l'ai dit précédemment, c'est le seul cas où l'orthographe note la position de l'accent. Ces facteurs peuvent être considérés comme indiquant le caractère relativement "marqué" de ce patron accentuel. Les sujets auraient dès lors évité ce patron parce qu'il leur aurait semblé moins naturel. Un facteur à ne pas négliger dans le cadre précis de cette étude est l'influence de l'écrit. En effet, pour faciliter la tâche du sujet, les mots-réponses étaient transcrits avant d'être lus. Si le mot-réponse devait être prononcé avec un accent final, le sujet aurait dû ajouter un accent sur la voyelle finale. Etant donné qu'aucun mot-test ne comportait d'accent graphique, les sujets ont peut-être considéré que cela aurait apporté une modification inopportune. En quelque sorte, le patron oxyton n'aurait jamais été choisi tout simplement parce que cela ne faisait pas réellement partie des choix possibles, introduisant de ce fait un biais expérimental assez important. Cette vision des choses me semble cependant indûment trop pessimiste. N'oublions pas que même si le but réel de l'expérience était dissimulé aux sujets, ils savaient que nous étions intéressées par la prononciation (ce qui justifie en même temps l'utilisation d'un enregistreur). Aux sujets les plus curieux, on a répondu que l'étude portait sur les accents régionaux. Pourquoi dès lors les sujets se seraient-ils tant fixés sur l'orthographe? Par ailleurs, certains sujets parlaient à voix haute pendant qu'ils écrivaient; si leur premier réflexe avait été de prononcer un mot oxyton, soit ils auraient ajouté un accent graphique, soit ils auraient manifesté leur indécision et posé une question. Or ce ne fut jamais le cas. Selon moi, penser que les sujets ne *pouvaient* pas déplacer l'accent avec la syllabe accentuée du mot-test, dans le cas des bisyllabes et des trisyllabes paroxytons, parce que cela aurait donné un patron oxyton, ce qui dans la phase écrite aurait nécessité une modification orthographique supplémentaire, n'est pas une interprétation entièrement satisfaisante. D'ailleurs, il existe bien un jeu de langage italien (cf. Bertinetto 1987) où les mots sont transformés de la façon suivante:

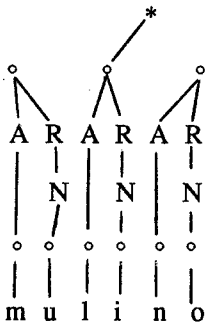
mano > *ma*gasà-*no*gosò
peli > *pe*ghesè-*li*ghisì
lunatici > *lu*gusù-*na*gasà-*ti*ghisì-*ci*ghisì

Le fait intéressant pour nous dans ce jeu, c'est la transformation du patron accentuel. En effet, il semble que chaque syllabe du mot, en l'occurrence paroxyton, donne naissance à un nouveau mot qui possède toujours le patron oxyton, indépendamment de tout le reste. Cela montre d'une part, qu'une modification du patron accentuel qui induit également une modification orthographique n'est pas une chose impossible pour les locuteurs de l'italien; et d'autre part, que l'on peut manipuler indépendamment les syllabes et le patron accentuel.

Finalement, on peut se demander pourquoi les sujets n'ont pas conservé le même patron accentuel que le mot-test dans tous les cas, puisque les manipulations demandées sont censées fonctionner dans un jeu de langage où la communication entre initiés doit être maintenue. L'explication est simple: les sujets ne sont pas "locuteurs natifs" de ce jeu de langage; de plus ils n'ont pas été placés dans la situation de devoir l'utiliser activement lors d'une conversation. En pratique, lors de cette expérience, on a surtout insisté sur l'inversion de syllabes, et s'est posée alors la question de la place de l'accent. Mais l'aspect langue secrète a sans doute été totalement oublié. On pourrait imaginer une expérience où l'on apprendrait aux sujets à inverser les syllabes tout en maintenant le patron accentuel, et dans un deuxième temps leur demander de faire l'inverse, à savoir leur présenter des mots en langue secrète, leur tâche étant alors de retrouver le mot italien d'origine. Cette expérience se ferait entièrement par oral puisque cette fois le choix du patron accentuel serait imposé (celui du mot-test). Si les sujets parviennent à effectuer cette expérience sans trop de difficulté, nous aurions un élément de plus en faveur de l'hypothèse selon laquelle le patron accentuel est une propriété du mot.

Quoi qu'il en soit, les résultats de cette étude montrent assez clairement qu'il est possible de choisir un patron accentuel et de le "coller" à un mot nouveau qui, pour le sujet, n'a aucun sens. La nette préférence

pour le patron paroxyton pourrait être le reflet de la proportion majoritaire de ce patron dans la langue[3]. Cette possibilité montre une certaine autonomie du patron accentuel par rapport au reste de la substance phonétique du mot, en ce sens que l'on peut agir indépendamment sur l'une ou sur l'autre composante, chacune faisant bien entendu partie intégrante de l'identité du mot. Ce genre de réalité peut être représentée à l'aide de la notion de "prosodie" au sens de l'école de Londres (Robins 1957), qui a été récemment rebaptisée "autosegment" par Goldsmith (1976) et suivants. Remarquons que pour l'école de Londres, l'accent est une prosodie de la syllabe (Robins, 1957: 6). Je préfère, pour ma part, distinguer à la manière de Garde (1968) l'unité accentuable, à savoir celle qui porte les manifestations phonétiques de l'accent, en l'occurrence la syllabe, et l'unité accentuelle, à savoir le domaine à l'intérieur duquel le contraste accentuel est effectué, en l'occurrence le mot. Pour représenter cette structure complexe, j'adopte un modèle dit "tridimensionnel" du type proposé par Encrevé (1988), auquel j'ajoute un squelette de positions syllabiques. Ce modèle à deux squelettes est également utilisé par Didier Demolin pour la représentation des tons en mangbetu. L'accent est donc conçu comme un autosegment qui s'ancre dans le deuxième squelette. Grâce à l'information morphologique, quelques principes et paramètres de portée plus ou moins étendue suffisent à rendre compte des divers patrons accentuels observés. Je ne dispose pas ici de la place nécessaire à l'exposition des modalités d'ancrage de l'accent, qui feront l'objet d'une publication ultérieure. Je me bornerai à fournir, à titre d'exemple, la représentation tridimensionnelle d'un mot paroxyton tel qu'il est réalisé dans la chaîne parlée d'un énoncé.



[3]Il est regrettable que nous ne disposions, à ce jour, d'aucun relevé systématique, effectué sur tout le lexique, des proportions relatives des trois patrons accentuels de l'italien. Je ne peux que citer les valeurs indicatives de Muljačić (1969: 491) obtenues à partir d'un texte de 10.000 formes: 60% de paroxytons, 4% de proparoxytons, et 3% d'oxytons. Les pourcents restants sont soit des clitiques inaccentués, soit des formes comprenant des clitiques, et qui ont été comptées comme portant l'accent sur une syllabe antérieure à l'antépénultième.

RÉFÉRENCES

Bertinetto, P.M. (1987) "Lingue segrete, segreti delle lingue. Alcuni problemi di fonologia italiana studiati attraverso un gioco linguistico", *Annali della Scuola Normale superiore di Pisa* 17, 889-920.

Bertinetto, P.M. (1988) "Felicity and Poverty of Experimental Phonology", *Quaderni del laboratorio di linguistica* 2, 85-111. Scuola Normale Superiore di Pisa. Version abrégée dans les *Proceedings of the 6th Int. Phonology Meeting*, Krems 1988.

Demolin, D. (1991) "L'analyse des segments, de la syllabe et des tons dans un jeu de langage Mangbetu", *Langages* 101, 30-50.

Demolin, D. (1992) *Le mangbetu: étude phonétique et phonologique*, thèse de doctorat, Université Libre de Bruxelles.

Encrevé, P. (1988) *La liaison avec et sans enchaînement. Phonologie tridimensionnelle et usages du fançais*, Paris, Seuil.

Garde, P. (1968) *L'accent*, Paris, Presses Universitaires de France.

Goldsmith, J. (1976) *Autosegmental Phonology*, Doctoral dissertation, M.I.T.

Hombert, J.-M. (1986) "Word Games: Some Implications for Analysis of Tones and Other Phonological Construct", in J. J. Ohala & J. J. Jaeger (eds.) (1986) *Experimental phonology*, Orlando, Fl.,Academic Press.

Muljačić, Ž. (1969) *Fonologia Generale e fonologia della lingua italiana*, Bologna, Il Mulino.

Ohala, J.J., (1984) "Explanation in Phonology : Opinions and Examples", in W.U. Dressler (ed.), *Phonologica*, Cambridge, Cambridge University Press.

Ohala, J.J. (1986) "Consumer's guide to evidence in phonology", *Phonology Yearbook* 3, 3-26.

Ohala, J. J., & Jaeger, J.J. (1986) *Experimental phonology*, Orlando, Fl., Academic Press.

Robins, R. H. (1957/1970) "Aspects of prosodic analysis", in *Diversions of Bloomsbury. Selected writings on linguistics*, Amsterdam, North-Holland, 1970. D'abord publié dans les *Proceedings of the University of Durham Philosophical Society*, Vol 1, Ser. B (Arts), number 1 (1957) 1-11.

Siegel, S. (1956) *Nonparametric statistics: For the Behavioral Sciences*, Tokyo, McGraw-Hill Kogakusha, Ltd.

ANNEXE

Tableau 1

Nombre de réponses par sujet et par type.
Légende: par = paroxyton, prop = proparoxyton, pour le reste de la légende voir texte.

Sujets	M	D	mD	dD	M par	M prop	D par	D prop	mD par	mD prop	dD par	dD prop	Tot par	Tot prop
1	50	40	15	25	23	27	15	25	15	0	0	25	38	52
2	48	42	42	0	48	0	42	0	42	0	0	0	90	0
3	30	60	42	18	30	0	42	0	42	0	0	18	76	18
4	36	54	39	15	33	3	39	15	39	0	0	15	72	18
5	37	53	42	11	37	0	42	11	42	0	0	11	79	11
6	41	49	48	1	41	0	48	1	48	0	0	1	89	1
7	31	59	42	17	31	0	42	17	42	0	0	17	73	17
8	37	53	42	11	37	0	42	11	42	0	0	11	79	11
9	42	48	42	6	42	0	42	6	42	0	0	6	84	6
10	24	66	22	44	4	20	22	44	22	0	0	44	26	64
11	77	13	1	12	36	41	1	12	1	0	0	12	37	53
12	47	43	42	1	47	0	42	1	42	0	0	1	89	1

19es J.E.P. - Bruxelles, 19 au 22 mai 1992.

EMERGENCE DE STRATEGIES OPPORTUNISTES DANS LA PROSODIE DE LECTURE : DEFINITION ET CARACTERISATION

GENEVIEVE CAELEN-HAUMONT

ICP / INPG, CNRS URA n° 368,
UNIVERSITE STENDHAL,
46 avenue F. VIALLET 38031 GRENOBLE CEDEX FRANCE

Résumé

Dans le cadre d'un travail général sur les relations entre prosodie et linguistique, nous avons dressé à propos d'un texte, un bilan des relations entre syntaxe, sémantique, pragmatique et les divers paramètres prosodiques, Fo, énergie, durée. Pour la plupart originaux, 6 modèles linguistiques et 24 indices prosodiques ont été définis. Modèles entre eux et d'autre part indices entre eux sont conçus comme concurrents.

Dans le cadre limité de cet article, nous nous proposons de définir et d'illustrer les stratégies d'association d'un modèle et d'un indice chez les locuteurs, en montrant en particulier le caractère opportuniste de ces stratégies dans leurs arguments syntaxiques, sémantiques, pragmatiques et psycho-cognitifs.

INTRODUCTION

Dans le domaine des relations entre linguistique et prosodie, les recherches ont abordé le problème sous l'angle de la phonétique, de la phonologie, de la syntaxe et de la sémantique. En ce qui concerne les aspects syntaxiques et sémantiques, diverses approches sont à prendre en considération. L'aspect syntaxique a été très tôt étudié puisqu'en 1890 déjà selon Hazaël-Massieux [1974], Sweet écrivait des propos encore très actuels : "of the two simple tones, the rising (and the level) is interrogative or expectant (suspensive), the falling affirmative or conclusive". Depuis, beaucoup de travaux ont été entrepris, surtout depuis le début des années 70.

L'approche est empirique ou théorique. Dans la perspective empirique, les analyses tendent à accumuler les données numériques (essentiellement de Fo et de la durée) pour décrire les constituants morpho-syntaxiques de la phrase, mots lexicaux et grammaticaux, groupes, phrases en fonction des contextes, des localisations dans les structures, des sous-unités ou des unités englobantes. Citons de manière très minimale les travaux de Lehiste [1970], de Fromkin [1980], Butterworth [1980] pour l'anglais. Plus récemment, c'est aussi la perspective d'Aubergé [1991] pour le français, ou de Fant [1991] pour le suédois. Tous ces travaux sont susceptibles d'apporter des informations précieuses aux systèmes de reconnaissance et à la synthèse.

Dans la perspective théorique, on recense aussi beaucoup de travaux, à la suite des génératives américains. Pour nous limiter au français, les principaux travaux ont été effectués par Di Cristo [1975, 1981], et dans une première version de l'analyse, par Martin [1975].

Dans le domaine de la sémantique, l'impulsion est donnée par les linguistes du Cercle de Prague [Karcevskij, 1931; Mathesius, 1939] qui introduisent les notions de thème (support) et de rhème (apport de l'information). Halliday [1967] reprenant en particulier cette conception, l'applique plus précisément à la prosodie et impulse l'idée de focus : "what is focal is 'new' information". Cette idée reprise par Bolinger [1972] a produit véritblement ses effets à partir du début des années 80, ce qui n'a pas manqué de provoquer d'ailleurs quelques débats passionnés entre les partisans de la suprématie de la syntaxe en matière de prosodie [Fromkin, 1983], et ceux de la suprématie de la sémantique [Cutler, 1983].

Ces idées convergent avec des travaux menés sur le français [Caelen-Haumont, 1978; Rossi 1981, 1985], sur l'anglais [Cutler, 1983; Horne, 1987; Fowler et Housum, 1987; Terken 1991], sur le suédois [Bruce, 1991].

A l'heure actuelle, on constate une véritable inflation des recherches sur les concepts de thème et de rhème en relation avec le focus. En comparaison avec l'extraordinaire variété des signifiés des énoncés, c'est curieusement le seul axe en sémantique vraiment développé.

C'est sur la base de ces constatations que mes recherches se sont développées en sémantique et pragmatique et

qu'il m'a semblé intéressant de tenter de dresser un bilan de ces relations sur les plans syntaxique, sémantique et pragmatique.

1. MODELES, INDICES, EXPERIMENTATION

L'hypothèse fondamentale qui sous-tend l'ensemble du travail est qu'en situation de communication orale, le traitement des contenus d'un énoncé et le traitement prosodique ne sont pas disjoints : il s'ensuit qu'il existe vraisemblablement une identité des structures profondes fondant l'oralisation d'un énoncé bien formé. Dans ces conditions, on peut admettre que cette identité est vérifiable numériquement. Dans cette perspective, 6 modèles linguistiques ont été définis tels qu'ils puissent prédire les niveaux de hauteur de Fo (et éventuellement des paramètres de la durée et de l'énergie) en des points-clé de l'énoncé, à savoir les mots lexicaux.

L'étude que nous avons menée porte donc sur les relations de coïncidences numériques entre 6 modèles prédictifs (2 syntaxiques, 3 sémantiques, 1 pragmatique) et les paramètres prosodiques. En ce qui concerne les indices prosodiques, outre les paramètres de l'énergie et de la durée, et les indices mélodiques "classiques" du maximum de Fo (ou FoM) et Fo moyen (ou Fom), nous introduisons un nouvel indice de Fo qui s'est révélé très efficace, à savoir la valeur absolue de l'écart de Fo (ou |ΔFo|) au sein du mot lexical.

Dans l'espace limité de cette communication, nous ne définirons les modèles utilisés que sous l'angle utilitaire de la compréhension de notre propos actuel. Nous renvoyons à d'autres communications pour de plus amples commentaires [1991a,b,c]. Les 6 modèles linguistiques se répartissent en 3 modèles qui proposent une analyse globale ou holistique de la structure de la phrase, et 3 modèles d'analyse locale des signifiés. Parmi les premiers, il existe un modèle d'analyse en constituants immédiats syntaxiques (HR), et deux modèles d'analyse en constituants immédiats sémantiques, les modèles de l'énonciation EN et ER reprennant l'organisation en thème / rhème mais dans une perspective d'analyse hiérarchique. Le modèle ER[1] se distingue du premier par une pondération plus importante accordée au constituant qui est porteur d'un "dire" (rhème) à propos d'un autre, support nécessaire à ce processus (thème).

Quant aux seconds, leur espace d'analyse se développe sur l'axe horizontal des relations lexicales, envisagées sous l'angle des relations de dépendance / indépendance syntaxiques (modèle DP), sous l'angle de la complexité sémantique intrinsèque et contextuelle (modèle CM), et enfin sous celui de la connaissance supposée (modèle CP) qui développe grandement le point de vue restreint de Prince [1983] sur les catégories de connaissances, à savoir celles déjà évoquées, inférées, nouvelles.

[1] Dans cette communication au niveau des résultats, nous ne distinguerons pas le modèle EN du modèle ER.

L'expérimentation a donc consisté à analyser les réalisations de 12 locuteurs (lecture d'un texte[2]) selon 3 consignes (1° lecture naturelle et intelligible 2° lecture très intelligible 3° lecture très très intelligible pour un ordinateur). Ces réalisations constituent une base de données pour laquelle on a posé manuellement environ 40 000 étiquettes empruntées à tous les niveaux d'analyse linguistique et prosodique. La confrontation entre les valeurs prédites par les 6 modèles linguistiques et les valeurs des 24 indices issus de Fo, de l'énergie et de la durée, nécessite bien entendu de neutraliser tous les effets de micro-mélodie, de variations intra- et inter-individuelles socio-linguistiques ou psychlogiques ... Pour ce faire, nous avons converti toutes les données numériques issues des valeurs prédictives des modèles, et des valeurs des indices prosodiques dans un espace à 4 niveaux. Cette méthode possède en outre l'avantage d'autoriser une comparaison fort intéressante de l'utilisation de l'espace prosodique chez chacun des locuteurs.

Précisons par ailleurs que tous les modèles entre eux et tous les indices d'un même paramètre sont concurrents entre eux, dans la recherche des scores les meilleurs de coïncidence.

2. LES STRATEGIES OPPORTUNISTES

L'observation des données numériques des réalisations des locuteurs, quelle que soit la consigne, nous a amenée à opérer d'emblée, tant les processus étaient distincts, une division d'une part entre les données qui sont issues de Fo, et d'autre part celles issues de l'énergie et de la durée. Pour ces derniers, les valeurs répondent clairement à une organisation croissante ou décroissante qui relève d'un mécanisme assez précis. Nous renvoyons à d'autres travaux pour plus d'informations [1991a] à ce sujet. Les choses sont très différentes pour les indices de Fo et c'est sur eux que repose le reste de notre propos.

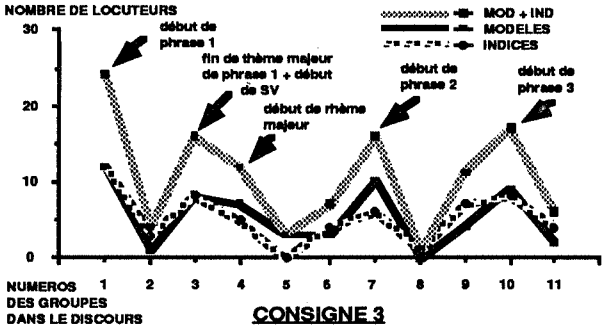
L'analyse des coïncidences entre les valeurs prédictives des 6 modèles et les valeurs numériques des 14 indices de Fo, a nécessité de manipuler 90 720 données par opération de tri. Au terme de cette manipulation et de l'analyse qui s'en est suivie, sont apparus des résultats validant les fondements de l'étude et faisant apparaître notamment des stratégies que les calculs de moyenage ont clairement établies. On a montré [1991a] par exemple que le taux de coïncidences moyen (sur 12 locuteurs) entre les valeurs prédictives d'un modèle et les valeurs d'un indice de Fo s'élevait de 81 à 85% en fonction de la consigne. Ces coïncidences analysées sur la base des

[2] le texte est le suivant : "D'éminents biologistes et d'éminents zoologistes américains ont créé pour des vers géants un nouveau phylum dans l'actuelle classification des nombreuses espèces vivantes. Ces longs vers prospèrent sur le plancher marin des zones sous-marines profondes. Des sources thermales chaudes y maintiennent une température moyenne élevée."

groupes pseudo-syntaxiques minimaux (GM) regroupent un nombre variable d'entre eux, soit en moyenne 6,8 mots lexicaux ou 2 à 3 GM. Ces coïncidences définissent donc des plages de stabilité pragmatique dans le discours (relation locuteur / énoncé) au cours desquelles les valeurs numériques d'un indice sont conformes aux valeurs prédites d'un modèle. A leur tour ces plages de stabilité s'organisent à l'analyse en stratégies significatives : c'est l'objet de cette communication.
Nous commenterons donc certains résultats exemplaires de ces stratégies et pour ce faire nous nous intéresserons à ces plages de stabilité. Par "stratégies opportunistes", nous entendons les moyens *hic et nunc* mis en oeuvre par le lecteur, attestant une organisation intelligente de son énoncé face à des sollicitations locales diverses telles que syntaxiques, sémantiques, pragmatiques et psycho-cognitives, sollicitations co-existantes dans tout énoncé et par là-même concurrentes.

2.1. ARGUMENTS SYNTAXIQUES

La situation de lecture impose au lecteur un texte dont les signifiés sont distribués en phrases et groupes syntaxiques de différents niveaux. Cet ensemble est organisé en vertu de lois grammaticales qui veillant à la cohérence de ce texte lui accordent un statut linguistique. Par exemple, en vertu de ces lois, il est recommandé pour la bonne compréhension de l'énoncé, que l'expression d'une nouvelle idée coïncide avec un début de phrase, et symétriquement, que la fin de phrase marque l'achèvement d'une première information, même provisoire. Sur le plan de la gestion de l'information, une phrase est un tout cohérent, une cellule de sens, une unité.



Graphique n° 1 : Points d'articulation majeurs syntaxiques et sémantiques de l'énoncé, tous locuteurs et toutes phrases confondus.

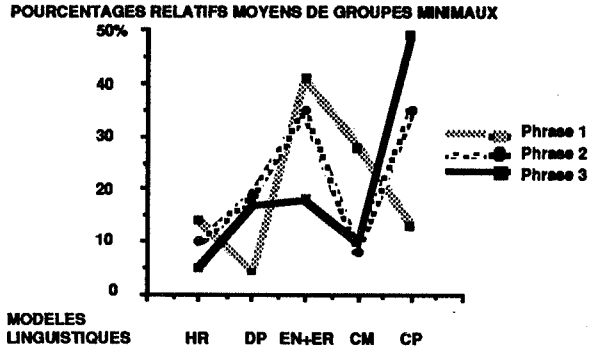
Et c'est vrai que les articulations syntaxiques majeures (en premier lieu les frontières de phrase, puis de groupes majeurs) constituent un argument favorable pour le changement de modèle et d'indice. Le graphique 1 ci-dessus, emprunté à la consigne 3, constitue un bon exemple de ce processus comme la courbe supérieure en grisé clair le montre bien.
Mais si à une phrase correspond l'expression d'une idée, on devrait s'attendre à ce que non seulement le début de phrase soit marqué par un nouveau modèle, mais que le modèle reste stable jusqu'à la fin de la phrase et change à nouveau pour le début de la suivante. Ceci, pour les phrases courtes. Pour neutraliser le paramètre de la longueur des phrases, nous admettrons pour les phrases longues, qu'il puisse y avoir sans remise en cause de l'homogénéité discursive, un changement de modèle pourvu qu'il soit opéré dans la même classe linguistique, par domaine, ou même pour placer l'hypothèse dans les cas les plus favorables de réalisation, par type (analytique ou holistique).
En fait le résultat moyen calculé sur l'ensemble des consignes est seulement de 37%. Dans les deux-tiers des cas, le locuteur semble donc être sensible à d'autres arguments.

2.2. ARGUMENTS SEMANTIQUES

Tout d'abord il faut bien voir que les constituants "syntaxiques" font également l'objet d'un autre découpage, sémantique cette fois-ci, et plus précisément énonciatif (thème / rhème), qui peuvent coïncider avec le découpage syntaxique. Lorsqu'ils ne coïncident pas, les débuts de rhème majeur sont tout autant les lieux du changement de modèle, comme le prouve également le graphique 1.
Mais un autre résultat est plus significatif encore. Pour garder une authenticité au caractère exemplaire de la démonstration, nous choisirons les illustrations parmi les résultas moyens des 12 locuteurs.
Sur le plan des signifiés, notre texte est caractérisé par le fait que la phrase 1 comporte le vocabulaire le plus spécialisé, la phrase 2, le vocabulaire le plus simple mais au contenu non prédictible (prospérité de vers géants dans les fonds sous-marins pourtant réputés généralement comme inhospitaliers), et la phrase 3, un vocabulaire un peu moins simple, mais fortement imprédictible (1° existence de sources thermales dans les océans 2° sources chaudes dans les bas-fonds réputés froids 3° température résultante élevée). Nous savons que la complexité des mots lexicaux est prise en charge par le modèle CM, les connaissances nouvelles et a fortiori inattendues, par le modèle CP[3].
Consultons le graphique n° 2 ci-dessous, en ne considérant pour le moment que les modèles CM et CP. Nous nous apercevons que le modèle CM en phrase 1, la plus complexe, vient en seconde position en ce qui concerne le nombre d'effectifs de groupes minimaux (derrière EN + ER), et qu'inversement il s'effondre à l'avant-dernière position, dans les phrases 2 et 3, les plus simples.
Par ailleurs si nous suivons l'évolution des effectifs en fonction du modèle CP, apte à traiter les signifiés inattendus, nous constatons qu'il est en mauvaise position en phrase 1 qui énonce un fait attendu, de par la spécialité et la qualité des agents mis en scène dès le début de cette phrase.

[3] Par simplification, nous traiterons du modèle pragmatique CP dans cette rubrique.

Inversement, dès la phrase 2, il se trouve en première position ex-aequo avec un modèle sémantique, et en phrase 3, il regroupe, loin devant les autres, le plus grand nombre d'effectifs.



Graphique n° 2 : distribution des effectifs des groupes minimaux en fonction des modèles et de chaque phrase, tous locuteurs confondus et toutes consignes confondues.

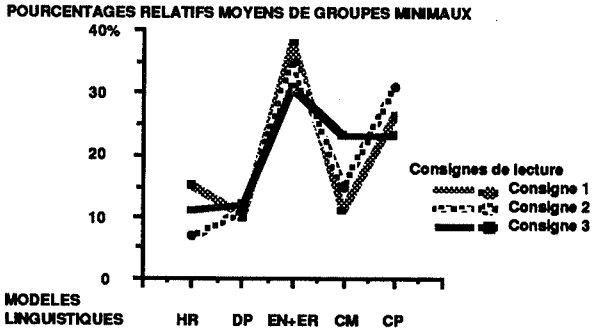
2.3. ARGUMENTS PRAGMATIQUES

Dans notre expérimentation la situation est représentée par l'introduction de consignes de lecture qui mettent en scène soit implicitement un récepteur humain potentiel (consignes 1 et 2) soit de manière explicite un ordinateur (consigne 3).

Nous envisagerons pour cette rubrique les faits sous l'angle des modèles et celui des indices.

Une première information concerne le débit (parole + pauses) : les résultats tous locuteurs et toutes phrases confondues montrent un net ralentissement de la consigne 1 (2.2 mots / seconde), à la consigne 2 (1.8 mots / seconde) et à la consigne 3 (1.05 mot / seconde).

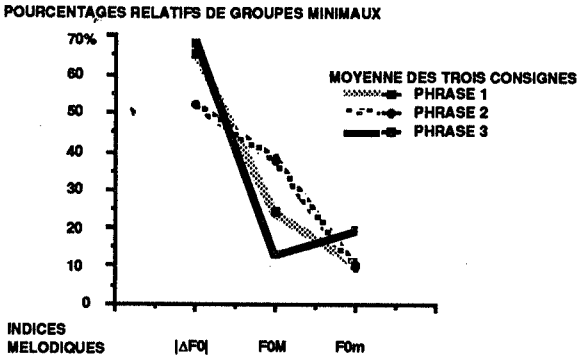
Par ailleurs en ce qui concerne les modèles, le graphique 3 ci-dessous présente la répartition des effectifs de groupes minimaux en fonction des différents modèles et selon les consignes.



Graphique n° 3 : distribution des effectifs des groupes minimaux en fonction des modèles et de chaque consigne, tous locuteurs confondus, toutes consignes confondues.

On remarque que la distribution des effectifs est cohérente d'une consigne à une autre. Mais les points de variation apportent une information intéressante. On observe en effet à propos des modèles les plus couramment utilisés (sémantiques, EN+ER, CM et pragmatiques CP), une évolution caractéristique des effectifs. Si quelle que soit la consigne, les modèles de l'énonciation sont prédominants, ils accusent cependant une perte d'effectifs de la consigne 1 à la consigne 3. Cette perte d'effectifs d'un modèle globaliste, profite en particulier à deux modèles analytiques, le modèle de la complexité CM et le modèle de la connaissance supposée CP. On remarque même que la progression du modèle CM en consigne 3, se fait également au dépens du modèle CP. Et le débit plus ralenti ou très ralenti s'accomode très bien de la plus grande faveur accordée aux modèles qui proposent une perspective analytique.

Les indices[4], quant à eux, révèlent le coût de l'oralisation, c'est-à-dire de la mise en oeuvre pragmatique et prosodique des signifiés textuels. Les faits sont clairement démontrés figure 4 ci-dessous. Il s'agit là des conditions pragmatiques de réalisation des énoncés.



Graphique n° 4 : distribution des effectifs des groupes minimaux en fonction des indices de Fo et de chaque consigne, tous locuteurs confondus, toutes consignes confondues.

Le premier fait est la suprématie évidente, quelle que soit donc la phrase ou la consigne, de l'indice que nous avons défini de manière originale, à savoir la valeur absolue de l'écart de Fo (ou |ΔF0|). Il apparaît cependant de manière caractéristique, que lorsque les conditions d'énonciation deviennent plus difficiles (enchaînement de la phrase 2 à la phrase 1, longue et au vocabulaire spécialisé), les locuteurs utilisent moins souvent l'indice |ΔF0|, au profit exclusif de l'indice du maximum de Fo (F0M). Lorsque les conditions deviennent plus faciles, la phrase 3 suivant la phrase 2, courte et au vocabulaire le plus simple, il se produit un retournement de la situation, et |ΔF0| obtient alors les effectifs les plus nombreux, supérieurs même à ceux de la phrase 1. Mais, fait intéressant, pour certains locuteurs, si les conditions d'énonciation sont facilitées en phrase 3, les

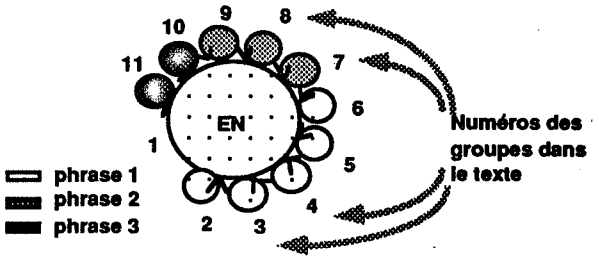
[4] Par simplification, les 14 indices ont été regroupés ici en 3 types fondamentaux.

conditions d'élocution deviennent inversement plus difficiles pour d'autres en fin de texte, et Fo moyen (F0m) croît également au dépens de F0M.
|ΔF0| est l'indice le plus précis mais aussi le plus coûteux dans la mesure où il nécessite de positionner dans la chaîne mélodique, des valeurs extrêmes absolues au sein du mot lexical, mais relatives au sein de l'énoncé de la phrase et du texte. Une lecture soignée est donc caractérisée par cet indice, mais lorsque les conditions deviennent plus difficiles, une des deux cibles disparaît (en l'occurrence le Fo minimum), puis dans les cas plus drastiques, les deux cibles, au profit de valeurs plus ou moins précises autour d'un seuil moyen positionnées sur un temps considérablement plus long. L'ensemble de ces comportements nous incite donc à penser que F0M et F0m sont en fait non des indices distincts, mais les formes progressivement détériorées de |ΔF0|, lorsque les conditions d'énonciation et/ou d'élocution deviennent plus difficiles.

2.4. ARGUMENTS PSYCHO-COGNITIFS

D'autres sollicitations existent encore qui poussent les lecteurs à sélectionner, de manière non consciente, certains modèles. Nous en connaissons au moins deux.
La première sollicitation d'ordre psycho-cognitif a déjà été entrevue dans les paragraphes précédents et il convient maintenant de l'expliciter entièrement. Elle est exprimée d'une part dans les méta-stratégies que l'on peut observer à propos de la bi-partition des modèles holistiques et des modèles analytiques, et d'autre part dans le relais qui est établi au cours de l'énoncé entre les modèles "fondateurs" et les modèles "expresseurs".
En effet les résultats (cf par exemple le graphique 2 ci-dessus), quelle que soit la consigne, nous montrent que les locuteurs dans notre expérimentation, ont très majoritairement recours aux modèles holistiques (EN, ER, HR) dans la première moitié de texte (phrase 1), alors que dans la deuxième moitié (phrases 2 et 3) les modèles analytiques dominent. Il est très intéressant d'un point de vue cognitif de savoir que ce processus se rencontre dans d'autres domaines, dans le domaine visuel par exemple, ou dans l'apprentissage. Toujours est-il que dans le cadre de notre texte, les modèles analytiques ont besoin, semble-t-il, que le cadre conceptuel et prosodique soit posé dans la première partie de l'énoncé pour se développer. Symétriquement, les premiers moments de la prise de parole sont sans doute les plus délicats à mettre en oeuvre dans la mesure où il faut créer ex nihilo à la fois les références prosodiques et les références conceptuelles de prise en charge pragmatique du texte par le locuteur, en fonction des contraintes qu'il identifie et des priorités qu'il accorde.
Les modèles holistiques proposent une structure plus simple, et en particulier les modèles de l'énonciation EN et ER très largement majoritaires, en proposant une partition binaire des éléments de signification selon un schéma très simple, laissent vraisemblablement moins de place à l'évaluation subjective. Ces modèles entrant vraisemblablement dans la compétence de tout locuteur usager de sa langue maternelle, sans doute aussi activés par les repères que procure l'image visuelle lors de la lecture, constituent certainement un recours commode pour le locuteur, dans certaines conditions délicates de mise en oeuvre du discours.
Mais parmi les modèles que l'on recense lorsque s'élabore le discours (dans notre expérimentation, la lecture), on trouve aussi le modèle CM, modèle de la complexité lexicale intrinsèque et contextuelle. Contrairement aux autres, ce n'est pas un modèle holistique, mais deux raisons peuvent être trouvées à son apparition en début de texte. Tout d'abord l'apprentissage de la signification intrinsèque et contextuelle des items lexicaux est un processus qui trouve son origine dès les premières semaines de la vie de l'homme, donc assimilé depuis longtemps par tout locuteur (les nôtres ont pour la plupart une formation supérieure), et dont les contenus sont donc rapidement disponibles en mémoire de travail lorsqu'il s'agit de communiquer les signifiés à l'auditeur potentiel.
La deuxième raison est contingente à notre texte. En effet la première phrase possède le registre de vocabulaire le plus difficile avec ses mots spécialisés et semi-spécialisés, et lorsque la contrainte sur l'intelligibilité —ou facilitation de l'intégration auditive et conceptuelle— est de plus en plus pressante, il est normal que l'utilisation d'un modèle qui gère la difficulté conceptuelle des mots lexicaux soit activée.
En conclusion donc, deux groupes de modèles émergent dans le cadre de notre étude, d'une part ceux qui apparaissent dès le début du texte, les modèles dits *"fondateurs"* de discours (EN, ER, HR, CM), essentiellement des modèles holistiques, mais nécessitant de toutes façons moins d'interprétation et d'évaluation, donc un traitement cognitif moins complexe et un accès plus direct, et d'autre part ceux qui interviennent ensuite, nécessitant que le cadre prosodique et lexical soient déjà posés, plus dépendants des contenus et de la situation de discours, et accordant plus d'attention aussi au destinataire, donc à tous égards moins impersonnels, et à ce titre recevant la qualification de modèles *"expresseurs"* (DP, CP). Nous pouvons alors interpréter maintenant complètement le graphique 3 déjà présenté, et comprendre que si les modèles holistiques (EN, ER, HR) restent prédominants quelle que soit la consigne, étant donné qu'ils sont une ressource commode dès qu'il existe une difficulté d'énonciation ou d'élocution, il n'empêche qu'ils sont tous globalement progressivement délaissés au profit des modèles plus analytiques, subjectifs et évaluatifs (CM, CP, DP).
Pour terminer avec les sollicitations psycho-cognitives, nous illustrerons les réalisations de la locutrice PE qui sont exemplaires de ce processus. En effet les réalisations de cette locutrice sont remarquables par le

fait qu'elle a utilisé un seul et même modèle pour l'organisation des valeurs mélodiques de son énoncé, à savoir le modèle EN. Fait plus surprenant encore, le taux des coïncidences entre les valeurs prédites par le modèle et les valeurs de Fo effectivement réalisées, est supérieur à la moyenne puisqu'il est de 90%.



LOCUTEUR PE : CONSIGNE 1

Graphique n° 5 : illustration d'une stratégie psycho-cognitive chez une locutrice.

Ainsi comme le montre la figure 5 ci-dessus, les arguments de sélection (non consciente) du modèle EN n'ont été ni syntaxiques, ni sémantiques, ni pragmatiques, mais psycho-cognitifs dans la mesure où ils identifient la planification d'une intention (non consciente) propre au locuteur, en définissant une stratégie subjective *d'ensemble* de l'énoncé par l'encodage homogène de la totalité des signifiés : et c'est justement par l'intermédiaire du modèle holistique EN à la structure simple, rapidement accessible, peu exigeant en investissement cognitif, qu'a pu se réaliser cette stratégie remarquable par son homogénéité.

CONCLUSION

En nous fondant sur notre expérimentation, nous avons montré que l'oralisation d'un texte dans la lecture avec consignes, est soumise à diverses contraintes, et pour nous limiter à la sphère linguistique, à des contraintes d'ordre non seulement syntaxique, sémantique, mais aussi prosodique, pragmatique et psycho-cognitif. Toutes ces contraintes sont en fait respectées dans l'énonciation d'un texte, mais le locuteur en fonction de ses motivations et de sa sensibilité particulière, va favoriser certaines d'entre elles : c'est ce qui constitue les coïncidences les plus fortes entre les valeurs prédictives des modèles et les valeurs effectivement réalisées dans l'énoncé. Indépendamment de ces coïncidences les plus fortes, existent également d'autres réseaux de coïncidences parallèles analysées par ailleurs [1991a], plus ou moins bien réalisés dans l'espace discursif, et aux taux moins élevés, mais ils n'en constituent pas moins aussi la trame fondamentale de l'énoncé.
Cette communication a montré également la légitimité des modèles linguistiques et des indices, modèles et indices souvent originaux, bien articulés les uns aux autres dans les divers énoncés. Il ressort de cette étude que les modèles syntaxiques sont largement insuffisants pour prédire les réalisations des locuteurs, mais aussi, étant donné l'immense variabilité des contenus de signification, que les modèles sémantiques, pragmatiques ou autres, sont loin d'avoir été tous conçus et appliqués.

REFERENCES

AUBERGE V. (1991), La synthèse de la parole : des règles au lexique, Thèse de doctorat de 3ème cycle, Université Mendès France, Grenoble.

BOLINGER D.L. (1972), Accent is predictible (if you're a mind reader), *Language*, 48, 633-644.

BRUCE G. (1991), The Exploitation of Pitch in Dialogue, Proc. of the 12th ICPhS, Aix-en-Provence, Vol. 1, 271-4.

BUTTERWORTH B. (1980), Some Constraints on Models of Language Production, in ed. Butterworth, *Language Production*, Vol 1, Speech and Talk, London : Academic Press, 423-459.

CAELEN-HAUMONT G. (1991a), Analyse des intéractions entre modèles syntaxiques, sémantiques, pragmatique et paramètres prosodiques : stratégies des locuteurs en réponse à des consignes de lecture de texte, thèse de doctorat d'état, Université de Provence, Aix-en-Provence.

CAELEN-HAUMONT G. (1991b), La valeur absolue du gradient de Fo : définition, localisation et distribution en lecture de texte français sous trois consignes différentes, Actes du 12ème ICPhS, Aix-en-Provence, Vol. 5, 182-5.

CAELEN-HAUMONT G. (1991c), Linguistic and Prosodic Features of Speaking Styles in French Texte Readings, Proceedings of ESCA Workshop, Barcelona, 14 1-7.

CUTLER A. (1983), Semantics, Syntax and Sentence Accent, Proc. Xth ICPhS II A, Utrecht : Foris Pub., 85-91.

DI CRISTO A. (1975), Recherches sur la structuration prosodique de la phrase française. Actes des 6èmes JEP, GALF-CNRS, Toulouse, 95-116.

FANT G. (1991), Units of Temporal Organization. Stress Groups versus Syllables and Words, Proc. of 12th ICPhS, Aix-en-Provence, Vol. 1, 247-250.

FOWLER C. A., HOUSUM J. (1987), Talker's Signalling of 'New" and 'Old' Words in Speech, and 'Listeners' perception and Use of the 'Distinction', *Journ. of Mem. and Lang.*, 26, 49, 489-504.

FROMKIN V. (1980), Errors in Linguistic Performance : Slips of the Tongue, Ear, Pen and Hand, in ed. V. A. Fromkin, New YorK : Academic Press.

FROMKIN V. (1983), The Independence and Dependence of Syntax, Semantics and Prosody, Proc. Xth ICPhS II A, Utrecht : Foris Publications, 93-97.

HALLIDAY M.A.K. (1967), Notes on Transitivity and Theme, II, *Journ. of Linguist.*, 3, 199-244.

HAZAEL-MASSIEUX M.-C. (1974), Situation et communication linguistique, Thèse de doctorat de 3ème cycle, Université de Paris III.

HORNE M. (1987b), Towards a Discourse-Based Model of English Sentence Intonation, *Working Papers*, 32, Departement of Linguistics, Lund University.

KARCEVSKIJ S. (1931), Sur la phonologie de la phrase, TCLP, 4, 188-227.

LEHISTE I. (1970), *Suprasegmentals*, MIT Press.

MARTIN P. (1975), Intonation et reconnaissance automatique de la structure syntaxique, 6èmes JEP, GALF, Toulouse, 52-62.

MATHESIUS V. (1939), O tak zvaném aktualnim cleneni vetrem, *Slovo a Slovesnost*, 5, 171-4.

PRINCE E.F. (1983), Toward a Taxinomy of Given-New Information, *Radical Pragmatics*, P. Cole ed., Academic Press, 223-255.

ROSSI M., DI CRISTO A., HIRST D., MARTIN P., NISHINUMA Y. (1981), *L'intonation, de l'acoustique à la sémantique*, Klincksieck, Paris.

ROSSI M. (1985), L'intonation et l'organisation de l'énoncé, *Phonetica*, 42, 135-153.

SWEET (1890), *Primer of Spoken English*, Oxford, Clarendon Press.

TERKEN J. M. B. (1991), Production and perception of Prosodic Prominence, Actes du 12ème ICPhS, Aix-en-Provence, Vol. 1, 288-293.

19es J.E.P. - Bruxelles, 19 au 22 mai 1992.

DUREE INTERSYLLABIQUE DANS LE GROUPE ACCENTUEL EN FRANCAIS

VALERIE PASDELOUP

**LABORATOIRE DE PSYCHOLOGIE EXPERIMENTALE,
UNIVERSITE LIBRE DE BRUXELLES, 117 AV. BUYL, 1050 BRUXELLES
ADRESSE PERMANENTE : 8 RUE DU CHATEAU-LANDON, 75010 PARIS**

Résumé

Une étude expérimentale réalisée à partir d'un corpus lu de 400 phrases (40 énoncés, 5 sujets, 2 répétitions), soit environ 6000 syllabes, analyse la durée intersyllabique dans le cadre du groupe accentuel. Trois facteurs sont pris en compte : la position de la syllabe inaccentuée par rapport à l'accent (pénultième, antépénultième etc...), le type d'accent du groupe accentuel (primaire ou secondaire) et le nombre de syllabes inaccentuées dont est composé le groupe accentuel. Cette étude met en évidence dans le groupe accentuel une configuration temporelle intersyllabique : un mouvement de ralentissement progressif du débit qui s'accentue sur la syllabe pénultième et qui culmine sur la syllabe accentuée. Ces résultats favorisent l'hypothèse d'une pré-programmation de l'accent ou de l'allongement final dès la 1ère syllabe du groupe accentuel.

1. INTRODUCTION

De nombreux travaux ont démontré le rôle de la pause et de l'allongement final dans la démarcation d'unités linguistiques (Goldman-Eisler, 1972 ; Grosjean et Deschamps, 1972 ; Duez, 1987). La configuration temporelle interne des unités ainsi démarquées apparaît, tout du moins en français, avoir suscité moins d'intérêt. Une des raisons principales qui semblent à l'origine de cette attitude est que le français a longtemps été considéré comme une langue à chronométrage syllabique ("syllable-timed") dont les syllabes brèves sont isochrones, et où la syllabe accentuée "se détache" du groupe de syllabes brèves inaccentuées qui la précède. Néanmoins certains travaux ont mis en évidence le fait que sur le plan temporel la syllabe accentuée est le résultat de la culmination d'un mouvement auquel participent tous les éléments du groupe accentuel: la durée des syllabes inaccentuées tend à progresser jusqu'à la syllabe accentuée dans un mouvement général de ralentissement progressif (Boudreault, 1970 ; Caelen, 1981 ; Pasdeloup, 1990a). L'évolution de la durée intersyllabique reste cependant encore peu étudiée en français.

2. METHODOLOGIE EXPERIMENTALE

La durée intersyllabique dans le groupe accentuel est analysée dans un corpus lu de 400 phrases (40 énoncés, 5 sujets, 2 répétitions), soit environ 6000 syllabes (Pasdeloup, 1990b). L'analyse des paramètres prosodiques (fréquence fonfamentale, durée et intensité) a permis d'extraire les indices à partir desquels est effectuée l'interprétation accentuelle. Deux expérimentations perceptives ont été réalisées sur une sélection de 16 phrases choisies parmi les 400 afin de tester la validité de ces indices. Tous les accents sont pris en compte. On distingue trois types d'accents : les accents primaires situés à la fin d'un mot et les accents secondaires qui ont une distribution différente (Fonagy, 1979 ; Verluyten, 1984 ; Hirst et Di Cristo, 1984 ; Rossi, 1985 ; Pasdeloup, 1988a, 1990b) : accents réalisés à l'initiale d'un mot (cas le plus fréquent), sur l'antépénultième d'un mot lexical et à la finale d'un morphème dans un mot polymorphémique ("anti/constitutionnel") ; les accents réalisés dans un mot monosyllabique constituent une catégorie particulière et sont nommés dans cet article accents monosyllabiques pour plus de facilités. L'accent primaire se caractérise sur le plan acoustique par un contour montant ou descendant de F0 de grande ou moyenne amplitude et par un allongement variable de la durée syllabique (généralement significatif). L'accent secondaire se caractérise par un contour montant de F0 de moyenne amplitude et par un très faible allongement de la durée syllabique (généralement non significatif).

Les syllabes sont étudiées dans le cadre du groupe accentuel auquel elles appartiennent. En français, un groupe accentuel est composé d'une syllabe accentuée généralement précédée de quelques syllabes inaccentuées. Le groupe accentuel est défini comme primaire s'il est constitué d'un accent primaire, comme secondaire s'il est constitué d'un accent secondaire etc... Trois facteurs sont pris en compte dans l'étude de la durée intersyllabique dans le groupe accentuel : la position de la syllabe inaccentuée par rapport à l'accent (pénultième, antépénultième etc...), le type d'accent du

groupe accentuel (primaire, secondaire ou monosyllabique) et le nombre de syllabes inaccentuées dont est composé le groupe accentuel.

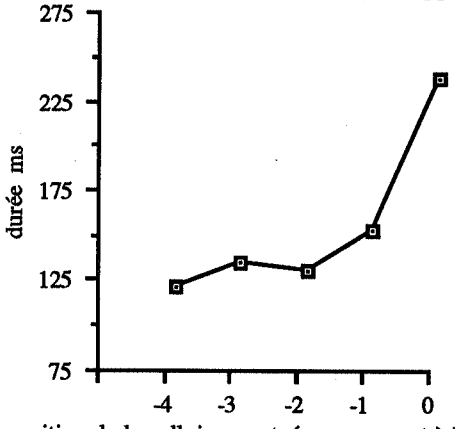
3. RESULTATS

3.1. Durée des syllabes inaccentuées dans le groupe accentuel en fonction de leur position par rapport à l'accent

Dans cette première analyse des résultats, on ne distingue pas le nombre de syllabes du groupe accentuel. On prend en compte la position de la syllabe inaccentuée dans le groupe accentuel par rapport à l'accent ; le décompte s'effectuant à partir de l'accent, de la fin du groupe accentuel à son début, est négatif ; la syllabe inaccentuée pénultième qui précède l'accent est notée e-1 (e pour élément), la syllabe antépénultième e-2 et ainsi de suite e-3, e-4 etc... Lorsque la syllabe inaccentuée fait partie d'un groupe accentuel constitué d'un accent primaire, elle est noté p-1, p-2, p-3 etc... ; si l'accent est un accent secondaire ou monosyllabique, la syllabe inaccentuée est notée respectivement s-1, s-2... et m-1, m-2... La syllabe accentuée est notée e0, et p0, s0 ou m0 s'il s'agit d'un accent primaire, secondaire ou monosyllabique.

3.1.1 Durée intersyllabique dans le groupe accentuel sans prise en compte du type d'accent

Figure 1 : Durée des syllabes dans le groupe accentuel selon leur position par rapport à l'accent

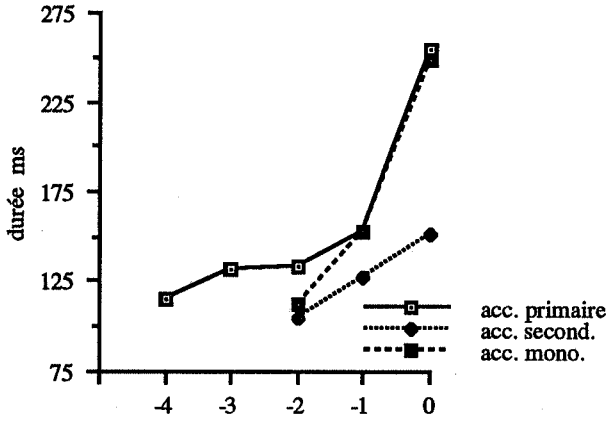


La figure 1 fait apparaître que la durée de la syllabe inaccentuée tend à augmenter dans le groupe accentuel au fur et à mesure qu'elle se rapproche de l'accent ; la syllabe pénultième est la plus marquée par ce phénomène de ralentissement, puisque sa durée augmente de 20% par rapport à la moyenne des syllabes inaccentuées qui la précède. Cependant, ce ralentissement n'est pas progressif : la durée de e-3 est légèrement supérieure à celle de e-2 (129ms contre 124ms), mais cette différence n'est pas significative. Selon les tests de Fisher PLSD et Scheffe F-test, les comparaisons des groupes de données suivants sont significatives à 95% : e0 vs. e-1, e0 vs. e-2, e0 vs. e-3, e0 vs. e-4, e-1 vs. e-2, e-1 vs e-3, e-1 vs. e-4, e-3 vs. e-4 (test de Fisher seulement) ; les comparaisons entre les groupes de données suivantes ne sont pas significatives : e-2 vs. e-3, e-2 vs. e-4.

3.1.2 Durée intersyllabique dans le groupe accentuel constitué d'un accent primaire, secondaire ou monosyllabique

La figure 2 fait apparaître un mouvement de ralentissement progressif du débit à l'intérieur des groupes accentuels primaire, secondaire et monosyllabique qui se renforce sur la syllabe pénultième et qui culmine sur la syllabe accentuée (cf. fig. 2). Dans le groupe accentuel primaire, la durée moyenne de la syllabe inaccentuée passe de 115ms en position -4 par rapport à l'accent, à 131ms en position -3, à 133ms en position -2 (antépénultième) et à 153ms en position -1 (pénultième) ; dans le groupe accentuel secondaire, la durée de la syllabe inaccentuée est de 105ms en position -2 et de 127ms en position -1 ; dans le groupe accentuel monosyllabique, la durée de la syllabe inaccentuée est de 113ms en position -2 et de 152ms en position -1. La syllabe pénultième est parmi les syllabes inaccentuées celle où le ralentissement est le plus fort : +21%, +21% et +35% d'allongement par rapport à la moyenne des syllabes inaccentuées qui précèdent, respectivement pour les groupes accentuels primaire, secondaire et monosyllabique.

Figure 2 : Durée des syllabes dans les groupes accentuels primaire, secondaire et monosyllabique



Dans le groupe accentuel primaire, les comparaisons des groupes de données suivants sont significatives (selon les tests de Fisher PLSD et Scheffe F-test) : p0 vs. p-1, p0 vs. p-2, p0 vs. p-3, p0 vs. p-4, p-1 vs. p-2, p-1 vs. p-3, p-1 vs. p-4, p-2 vs. p-4 (test de Fisher seulement), p-3 vs. p-4 (test de Fisher seulement) ; les comparaisons entre les groupes de données suivantes ne sont pas significatives : p-2 vs. p-3. Dans le groupe accentuel secondaire, les

comparaisons entre tous les groupes de données sont significatives : s0 vs. s-1 (test de Fisher seulement), s0 vs. s-2, s-1 vs. s-2 (test de Fisher seulement). Dans le groupe accentuel monosyllabique, les comparaisons entre tous les groupes de données sont significatives : m0 vs. m-1, m0 vs. m-2, m-1 vs. m-2.

Cette première série de résultats met en évidence une configuration globale de ralentissement de même type dans les groupes accentuels primaires, secondaires et monosyllabiques. La prise en compte du type d'accent qui constitue le groupe accentuel est néanmoins essentiel puisque la configuration de ralentissement des groupes accentuels secondaires ne se superpose pas aux configurations de ralentissement des groupes accentuels primaires et monosyllabiques qui elles sont proches (cf. fig. 2). A position égale de la syllabe par rapport à l'accent, les durées des syllabes dans les groupes accentuels secondaires sont inférieures à celles des syllabes dans les groupes accentuels primaires et monosyllabiques : la durée moyenne de s0 (syllabe portant l'accent secondaire) est ainsi proche de celle de p-1 et m-1 (151ms contre 153 et 152ms), la durée moyenne de s-1 est proche de celle de p-3 (127ms contre 131ms) et la durée moyenne de s-2 est proche de celle de p-4 et m-2 (105ms contre 115ms et 113ms). L'accent secondaire, contrairement aux accents primaire et monosyllabique se caractérise principalement par une montée de F0 et par un faible allongement syllabique ; malgré la faible participation de l'indice de durée dans la réalisation de l'accent secondaire, on constate cependant une configuration de ralentissement dans le groupe accentuel secondaire.

3.2. Durée des syllabes inaccentuées en fonction du nombre de syllabes du groupe accentuel

Dans cette analyse, on prend en compte le nombre de syllabes qui constituent le groupe accentuel et, comme précédemment, la position de la syllabe inaccentuée par rapport à l'accent. On regroupe les groupes accentuels en fonction de leur nombre de syllabes inaccentuées : groupes accentuels comprenant 1 syllabe inaccentuée, 2 syllabes inaccentuées etc... Les groupes accentuels qui comprennent 1 et 2 syllabes inaccentuées représentent 75% des cas et les groupes accentuels qui comprennent de 1 à 4 syllabes inaccentuées représentent 96% des cas ; ces résultats confirment d'autres travaux relatifs à la longueur du groupe accentuel en français (Boudreault, 1968 ; Wenk et Wioland, 1982).

3.2.1. Durée intersyllabique dans le groupe accentuel sans prise en compte du type d'accent

Plus le nombre de syllabes inaccentuées du groupe accentuel augmente, plus la durée totale du groupe augmente (cf. tableau 1) ; cette observation confirme l'hypothèse d'un principe de chronométrage syllabique en français. La durée totale du groupe accentuel n'augmente pas de façon linéaire avec le nombre de syllabes inaccentuées du groupe accentuel : la durée moyenne des syllabes inaccentuées tend plus ou moins à diminuer quand le nombre de syllabes du groupe accentuel augmente ; il ne s'agit pas véritablement d'un phénomène marqué de chronométrage accentuel, mais seulement d'une légère tendance, puisque la durée moyenne de la syllabe inaccentuée décroît irrégu-

Tableau 1 : Durée moyenne des syllabes dans le groupe accentuel en fonction de sa longueur (les groupes accentuels composés de 5 syll. inacc. ne représentent que 2% des cas (30 cas))

nombre de syllabes inaccentuées dans le groupe accentuel	1	2	3	4	5
durée totale du groupe accentuel	377ms	472ms	663ms	765ms	842ms
durée des syllabes inaccentuées	144ms	132ms	137ms	131ms	126ms
durée de la syllabe pénultième	144ms	148ms	147ms	150ms	146ms

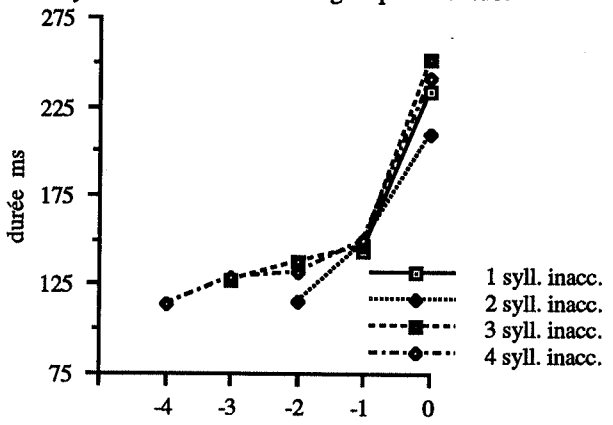
Tableau 2 : Durée moyenne des syllabes dans le groupe accentuel primaire en fonction de sa longueur (les groupes accentuels composés de 5 syll. inacc. ne représentent que 2.5% des cas (25 cas))

nombre de syllabes inaccentuées dans le groupe accentuel primaire	1	2	3	4	5
durée totale du groupe accentuel	395ms	543ms	686ms	766ms	862ms
durée des syllabes inaccentuées	146ms	147ms	141ms	131ms	128ms
durée de la syllabe pénultième	146ms	161ms	151ms	151ms	154ms

lièrement et peu dans les groupes accentuels composés de 2 à 5 syllabes inaccentuées. La durée moyenne de la syllabe pénultième est relativement stable quelque soit la longueur du groupe accentuel.

Pour les groupes accentuels qui comprennent 1, 2, 3 et 4 syllabes inaccentuées, qui représentent 96% des réalisations, on remarque dans tous les cas une configuration de ralentissement (cf. fig. 3). Dans les groupes accentuels qui comprennent 5 syllabes inaccentuées (2% des cas, 30 cas), on observe seulement une tendance au ralentissement.

Figure 3 : Durée des syllabes en fonction du nombre de syllabes inaccentuées du groupe accentuel



3.2.2 Durée intersyllabique dans le groupe accentuel primaire, secondaire et monosyllabique

Les groupes accentuels primaires sont composés en moyenne d'un plus grand nombre de syllabes que les groupes accentuels secondaires et monosyllabiques. Les groupes primaires qui comprennent de 1 à 4 syllabes inaccentuées représentent 95% des cas (les groupes qui comprennent 1 et 2 syllabes inaccentuées représentent 60% des cas). Les groupes secondaires qui comprennent 1 ou 2 syllabes inaccentuées représentent 89% des cas. Les groupes monosyllabiques qui comprennent 1 ou 2 syllabes inaccentuées représentent 96% des cas.

Tableau 3 : Durée moyenne des syllabes dans le groupe accentuel secondaire en fonction de sa longueur
(les groupes accentuels composés de 3 syll. inacc. ne représentent que 5% des cas (22 cas))

nombre de syllabes inaccen. dans le groupe accentuel second.	1	2	3
durée totale du groupe accentuel	260ms	390ms	532ms
durée des syllabes inaccentuées	116ms	117ms	123ms
durée de la syllabe pénultième	116ms	134ms	125ms

Tableau 4 : Durée moyenne des syllabes dans le groupe accentuel monosyllabique en fonction de sa longueur

nombre de syllabes inaccentuées dans le groupe accentuel monosyl.	1	2
durée totale du groupe accentuel	408ms	483ms
durée des syllabes inaccentuées	155ms	127ms
durée de la syllabe pénultième	155ms	145ms

Quelque soit le type de groupe accentuel, primaire, secondaire ou monosyllabique, on remarque un principe de chronométrage syllabique puisque plus le nombre de syllabes inaccentuées du groupe accentuel augmente plus la durée totale du groupe augmente (cf. tableaux 2, 3 et 4). On observe cependant des différences entre, d'une part, les groupes accentuels secondaires et, d'autre part, les groupes accentuels primaires et monosyllabiques : alors que dans ces derniers la durée moyenne des syllabes inaccentuées tend légèrement à diminuer quand le nombre de syllabes du groupe accentuel augmente, dans les groupes accentuels secondaires la durée moyenne des syllabes inaccentuées ne diminue pas quand le nombre de syllabes du groupe accentuel augmente. Pour les groupes accentuels primaires, la durée moyenne de la syllabe inaccentuée évolue de 146.5ms quand le groupe est composé de 1 ou de 2 syllabes inaccentuées, à 128ms quand il en est composé de 5 ; pour les groupes accentuels monosyllabiques, la durée moyenne de la syllabe inaccentuée est de 155ms quand le groupe est composé de 1 syllabe inaccentuée et de 127ms quand il en est composé de 2. La tendance au chronométrage accentuel ne concerne par conséquent que les groupes accentuels primaires et monosyllabiques.

Figure 4 : Durée des syllabes en fonction du nombre de syllabes inaccentuées du groupe accentuel primaire

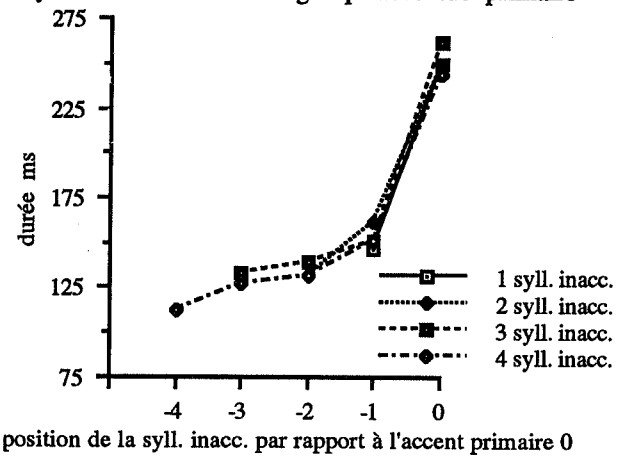


Figure 5 : Durée des syllabes en fonction du nombre de syllabes inaccentuées du le groupe accentuel secondaire

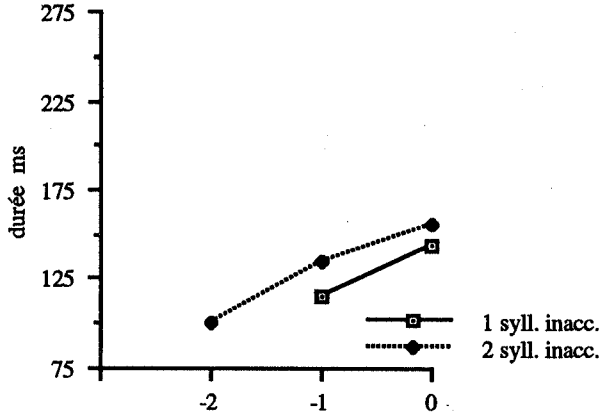
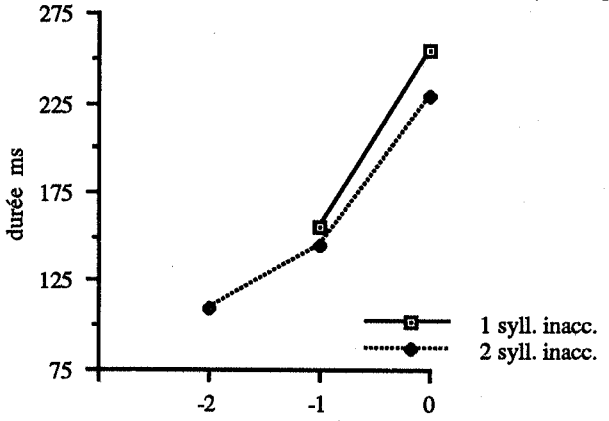


Figure 6 : Durée des syllabes en fonction du nombre de syllabes inaccentuées du groupe accentuel monosyllabique



Pour les groupes accentuels primaires composés de 1, 2, 3 et 4 syllabes inaccentuées et pour les groupes accentuels secondaires et monosyllabiques composés de 1 et 2 syllabes inaccentuées, on observe systématiquement un mouvement de ralentissement progressif du débit qui s'accélère sur la syllabe pénultième et culmine sur la syllabe finale (cf. fig. 4, 5 et 6). Ces groupes accentuels représentent la majorité des réalisations (pour les groupes accentuels primaires, secondaires et monosyllabiques respectivement 95%, 89% et 96%). Pour les autres groupes accentuels, les groupes accentuels primaires composées de 5 syllabes inaccentuées et les groupes accentuels secondaires composés de 3 syllabes inaccentuées, on remarque des configurations irrégulières de ralentissement ; ces résultats doivent être considérés avec précaution puisqu'ils ne représentent respectivement que 2.5% et 5% des cas.

4. CONCLUSION

La durée des syllabes inaccentuées en français n'est pas isochrone. Son étude dans le cadre du domaine phonologique du groupe accentuel met en évidence une configuration temporelle intersyllabique spécifique. Dans la majorité des cas, c'est-à-dire pour les groupes accentuels primaires qui comprennent 1, 2, 3 et 4 syllabes inaccentuées ainsi que pour les groupes accentuels secondaires et monosyllabiques qui comprennent 1 et 2 syllabes inaccentuées, on observe un mouvement de ralentissement progressif du débit qui s'accentue sur la syllabe pénultième et qui culmine sur la syllabe accentuée. Dans le cas même du groupe accentuel secondaire où l'allongement de la syllabe accentuée (d'environ 20%) ne joue qu'un rôle marginal dans sa réalisation, la durée des syllabes inaccentuées tend à progresser de façon significative vers la syllabe accentuée.

Ces résultats confirment des travaux antérieurs menés sur le français et le québecquois selon lesquels la durée des syllabes inaccentuées tend à progresser jusqu'à la syllabe accentuée (Boudreault, 1970 ; Caelen, 1981 ; Pasdeloup, 1988b, 1990a). Dans Pasdeloup (1990a), une étude menée sur un corpus de 96 phrases réitérées en syllabes ma-ma-ma met en évidence le fait que la phrase s'organise temporellement en phases de ralentissement de plus ou moins grande amplitude, chaque phase temporelle étant suivie d'une réinitialisation de la durée syllabique.

Les importantes variations intrinsèques de durée intersyllabique - liées aux différences segmentales entre les syllabes - masquent généralement, dans l'étude d'un énoncé unique, ces configurations de ralentissement du débit. L'étude d'une grande base de données (ici environ 6000 syllabes) permet de minimiser l'influence de ces variations intrinsèques de durée et de laisser apparaître des mouvements de ralentissement du débit dans le groupe accentuel.

Le mouvement de ralentissement du débit dans le groupe accentuel s'organise conjointement avec le principe de chronométrage syllabique : plus le nombre de syllabes inaccentuées du groupe accentuel augmente, plus la durée totale du groupe augmente. La durée totale des groupes accentuels primaires et monosyllabiques n'augmente pas néanmoins de façon linéaire avec leur nombre de syllabes : la durée moyenne des syllabes inaccentuées tend à diminuer lorsque le nombre de syllabes du groupe accentuel augmente (Wenk & Wioland, 1982). Par conséquent, bien que le principe de chronométrage syllabique soit prépondérant en français, ce principe semble modulé par une légère tendance au chronométrage accentuel.

Selon nous, il est peu probable que ces ralentissements du débit dans le groupe accentuel soient perçus et qu'ils

aient une fonction pré-indicatrice de l'accent. La syllabe pénultième est parmi les syllabes inaccentuées celle où le ralentissement est le plus marqué ; dans les groupes accentuels primaires et secondaires, l'allongement de la syllabe pénultième relativement aux syllabes inaccentuées qui la précèdent est proche du seuil différentiel de durée (Rossi, 1972) ; dans les groupes accentuels monosyllabiques, il dépasse cependant ce seuil puisque l'allongement moyen de la pénultième est de 35%. Nous émettons l'hypothèse que les ralentissements du débit dans le groupe accentuel ne sont pas perçus et qu'ils résulteraient de contraintes biologiques et cognitives liées à la programmation et à la production des syllabes accentuées. Benguerel et D'Arcy (1986) démontrent à partir d'une série de tests perceptifs que ce qui est perçu comme une séquence régulière de syllabes ne correspond généralement pas sur le plan acoustique à une séquence de syllabes isochrones, mais à une séquence de syllabes qui décélèrent ; ces auteurs émettent l'hypothèse que "due to articulatory, linguistic and other constraints, what is intended to be regular at the pre-production stage becomes time-warped, usually in the direction of deceleration, resulting in a lengthening which is most marked at the end of an utterance or of a breath-group"(p. 244).

En conclusion, la syllabe accentuée ne se "détache" pas du groupe de syllabes phonologiquement brèves qui la précède. L'analyse du phénomène d'allongement syllabique ne doit pas se limiter aux seules syllabes accentuées puisqu'ils affectent toutes les syllabes du groupe accentuel. Selon Butterworth et Goldman-Eisler (1979) et Fowler (1980), la structuration temporelle d'unités segmentales et suprasegmentales nous informe sur les mécanismes de programmation et de plannification. Dans cette optique, les résultats de cette étude expérimentale favorisent l'hypothèse que l'accent ou l'allongement de la syllabe finale du groupe accentuel serait pré-programmé dès la 1ère syllabe du groupe accentuel. Sternberg et al. (1988), dans une analyse du timing d'énoncés à débit rapide, émettent l'hypothèse que le groupe accentuel serait l'unité d'action dans la pré-plannification des durées d'un énoncé.

BIBLIOGRAPHIE

Benguerel, A.-P. ; D'Arcy, J. (1986) Time-warping and the perception of rhythm in speech, *Journal of Phonetics*, **14**, 231-246.

Boudreault, M. (1968) *Rythme et mélodie de la phrase parlée en France et au Québec*, Les Presses de l'Université Laval, Québec, 273 p.

Boudreault, M. (1970) Le rythme en langue franco-canadienne, in *Analyse des faits prosodiques, Studia Phonetica*, **3**, Didier.

Butterworth, B. ; Goldman-Eisler, F. (1979) Recent Studies on Cognitive Rhythm, in *Of Speech and Time*, A. W. Siegman, S. Feldstein eds., Lawrence Erlbaum Associates, New Jersey, 211-224.

Caelen, G. (1981) *Structures prosodiques de la phrase énonciative simple et étendue*, Thèse de 3ème cycle, Hamburger Phonetische Beiträge, Bd. 34, Buske, Hamburg.

Duez, D. (1987) *Contribution à l'étude de la structuration temporelle de la parole en français*, Thèse de Doctorat d'Etat, Université de Provence, Aix-Marseille 1.

Fónagy, I. (1979) L'accent français : accent probabilitaire, L'accent en français contemporain, *Studia Phonetica*, **15**, Didier, 123-233.

Fowler, C. (1980) Coarticulation and Theories of Timing Extrinsic, *Journal of Phonetics*, **8 (1)**, 113-133.

Goldman-Eisler, F. (1972) Pauses, clauses, sentences, *Lang. Speech*, **15**, 103-113.

Grosjean, F. ; Deschamps, A. (1972) Analyse des variables temporelles du français spontané, *Phonetica*, **26**, 129-156.

Hirst, D. J. ; Di Cristo, A. (1984) French intonation : A Parametric Approach, *Die Neueren Sprachen*, **83:5**, 554-569.

Pasdeloup, V. (1988a) Essai d'analyse du système accentuel du français : distribution de l'accent secondaire, *Actes des 17èmes Journées d'Etudes sur la Parole*, Nancy, 20-23 Septembre 1988, 65-70.

Pasdeloup, V. (1988b) Temporal phases in French : an acoustic study of reiterant speech, *Proc. of 7th Symposium of Federation of Acoustic Societies of Europe*, Edimbourg, 22-26 Août 1988, 1397-1404.

Pasdeloup, V. (1990a) Organisation de l'énoncé en phases temporelles : analyse d'un corpus de phrases réitérées, *Actes des18èmes Journées d'Etudes sur la Parole*, Montréal, 28-31 Mai 1990.

Pasdeloup, V. (1990b) Modèle de règles rythmiques du français appliqué à la synthèse de la parole, *thèse de doctorat nouveau régime*, Universitée d'Aix-Marseille I.

Rossi, M. (1972) Le seuil différentiel de durée, *Papers in memory of Pierre Delattre*, Mouton, La Hague, 435-450.

Rossi, M. (1985) L'intonation et l'organisation de l'énoncé, *Phonetica*, **42**, 135-153.

Sternberg, S. ; Knoll, R. L. ; Wright, C. E. (1988) Motor Programs and Hierarchical Organization in the Control of Rapid Speech, *Phonetica*, **45**, 175-197.

Verluyten, S. P. (1984) Phonetic Reality of Linguistic Structures : the Case of (Secondary) Stress in French, *Proc. of the Tenth International Congress of Phonetic Sciences*, Utrecht, M. P. R. Van den Brœcke, A. Cohen eds., 522-526.

Wenk, B. J. ; Wioland, F. (1982) Is French really syllable-timed ? *Journal of Phonetics*, **10**, 193-216.

19es J.E.P. - Bruxelles, 19 au 22 mai 1992.

L'INTERACTION DE LA PROSODIE AVEC LES VARIATIONS INTRINSEQUES DE FO DES VOYELLES EN ARABE

M. YEOU

INSTITUT DE PHONETIQUE (URA 1027)
UNIVERSITE LA SORBONNE NOUVELLE. 19 RUE DES BERNARDINS 75005 PARIS.

Résumé

ABSTRACT

The intrinsic fundamental frequency (IFo) difference between high and low vowels in Arabic was investigated in different prosodic contexts in which nuclear stress, vowel quantity, sentence position and contrastive stress were controlled. Findings, reported in this paper, confirm previous work on the interaction of prosodic factors with vowel intrinsic pitch differences and point to the necessity of separating Fo variations inherent to a segment from those required by prosody. This suggests that in implementing intonation in speech synthesis and automatic speech recognition, factoring out IFo effects is not a straightforwad task. Larger intrinsic differences were found in long vowels, in stressed vowels, in initial position and under contrastive stress.

INTRODUCTION

C'est un fait bien connu que les voyelles fermées ont un fondamental plus élevé que celui des voyelles ouvertes dans un contexte comparable. Ce phénomène a été observé dans de nombreuses langues et il tend a être considéré universel. De nombreux chercheurs ont proposé différentes explications physiologiques et acoustiques (voir pour une revue [1], [2] et [3]). Dans la plupart des études, l'étude s'effectue dans un contexte intonationel unique. Des recherches récentes ont souligné l'importance de l'interaction des facteurs prosodiques avec ce phénomène [4], [5] [6].

L'objetif de cette communication est d'étudier systématiquement l'effet des facteurs prosodiques comme l'accent de mot, la quantité vocalique, la position dans la phrase et l'accent emphatique sur les différences intrinsèques du fondamental entre les voyelles fermées et les voyelles ouvertes de l'arabe standard moderne.

METHODE D'ANALYSE

2.1 CORPUS ET LOCUTEURS

Pour étudier les différences intrinsèques des voyelles brèves et longues dans un contexte accentué et inaccentué, nous avons utilisé des logatomes trisyllabiques du type:

/lVl'VllVn/
/lV:l'V:lVn/
où V= /i, a, u/ et V: = /i:, a:, u:/
'V= voyelle accentuée

La voyelle choisie est la même dans chaque logatome. Ces logatomes, dont la structure phonotactique est bien formée, ont été introduits dans des phrases cadres du type "*qa:la ___ marratajni*" (Il a dit __ deux fois). Ce corpus a été enregistré par trois locuteurs marocains. Toutes les phrases ont été répétées 5 fois. Le nombre total des voyelles à étudier s'élève à 180 (5 répétitions x 6 voyelles accentuées x 6 voyelles inaccentuées x 3 locuteurs).

Afin d'étudier l'effet de la position dans la phrase sur FoI, nous avons retenu seulement [a:] et [i:] comme représentant les voyelles ouvertes et fermées. Ces deux voyelles ont été introduites dans des prénoms en position initiale et finale dans la même phrase de type déclarative:

ʒala:l ʔahda: dirhaman li-ʒali:l (Jalal a offert un dirham à Jalil)

ʒalil ʔhda: dirhaman li-ʒala:l (Jalil a offert un dirham à Jalal)
Le fait que ces mots propres soient interchangeables permet un contrôle systématique des variables comme le contexte segmental, le nombre de syllabes, la structure syntaxique et le contenu sémantique.

Ces mêmes phrases ont été utilisées pour étudier l'influence du focus sur les différences intrinsèques entre [a:] et [i:]. Pour assurer le placement propre du focus, les locuteurs avaient comme consigne de placer l'accent emphatique, par exemple, sur le premier mot de la phrase: ***ʒalaːl*** *ʔahda: dirhaman li-ʒalil !* Pour cela, nous leur avons demandé de produire des phrases qui montrent clairement que c'est, par exemple, Jalal qui a offert un dirham à Jalil. En outre, chaque phrase a été repétée en augmentant le degré de focus sur 5 niveaux, allant du moins emphatique au plus emphatique possible. Enfin, les deux mots emphatiques ont été placés à la position initiale et à la position finale de la phrase.

Deux locuteurs ont participé à l'enregistrement de ce corpus. Le total des phrases est le suivant:
Phrases déclaratives simples: 2 locuteurs x 2 voyelles x 2 positions x 5 répétitions = 40 phrases.
Phrases avec focus: 2 locuteurs x 2 voyelles x 5 niveaux de focus x 2 positions x 3 répétitions = 120 phrases.

2.2 MESURES

L'ensemble du corpus a été numérisé avec une fréquence d'échantillonage de 10 Khz et une dynamique de 16 bits, et analysé à l'aide d'une méthode temporelle d'extraction de Fo sur Signalyze, un logiciel de traitement du signal sur Macintosh (Keller, 1989). Nous avons mesuré la valeur maximale de la fréquence fondamentale dans la syllabe accentuée et la syllabe inaccentuée.

RESULTATS

3.1 L'EFFET DE L'ACCENT ET LA DURÉE

La figure n° 1 montre que la différence entre la fréquence fondamentale intrinsèque des voyelles fermées et ouvertes varie systématiquement en fonction de l'accent du mot et de la quantité vocalique.

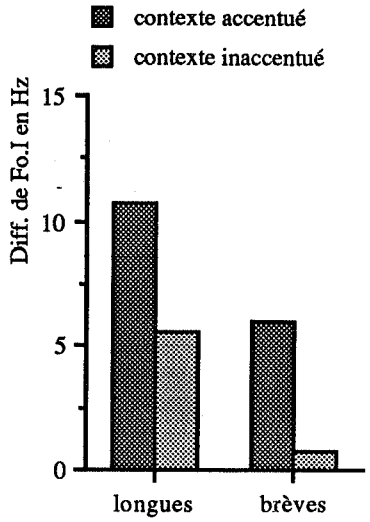


Fig 1: différences intrinsèques des voyelles en arabe en fonction de l'accent de mot et de la durée (3 locuteurs).

Les différences intrinsèques de Fo sont plus marquées entre les voyelles accentuées qu'entre les voyelles inaccentuées, et entre les voyelles longues qu'entre les voyelles brèves. Comme le montrent les tableaux n°1 et n°2 le rapport VF (voyelle fermée)/ VO (voyelle ouverte) s'élève dans l'ordre suivant:

i. voyelles brèves inaccentuées (0,46 %)
ii. voyelles brèves accentuées (3 %) et V. longues inaccentuées (3,3 %)
iii. voyelles longues accentuées (5,6 %).

	Voyelles inaccentuées			Voyelles accentuées		
	i	**u**	**a**	**i**	**u**	**a**
MY	141	140	140	164,8	163,8	157,2
LQ	171,4	170,2	169,4	190,8	189,8	184
AM	180	180	180	227,2	225,8	223,2
Σ Loc	164,3	163,4	163,1	194,2	193,1	188,1
i+u/a	**0,46 %**			**3 %**		
Δ Fo	**0,75 HZ**			**5,6 Hz**		

Tableau n° 1: Valeurs moyennes de Fo des voyelles brèves accentuées et inaccentuées (5 répétitions). ΔFo indique la valeur suivante: Fo de [i]+ Fo de [u] - Fo de [a].

	Voyelles inaccentuées			Voyelles accentuées		
	i:	**u:**	**a:**	**i:**	**u:**	**a:**
MY	158,4	157,2	149,6	195,4	196,2	178,4
LQ	179,4	177,4	173,8	192,6	187,2	184,2
AM	211,2	213,6	207,6	228,6	233,8	222,2
Σ Loc	183	182,7	177	205,5	205,7	194,9
i+u/a	**3,3 %**			**5,5 %**		
Δ Fo	**6 Hz**			**10,7 Hz**		

Tableau n° 2: Valeurs moyennes de Fo des voyelles brèves accentuées et inaccentuées (5 répétitions).

La comparaison des différences intrinsèques entre les voyelles brèves accentuées de l'arabe avec celles d'autres langues (voir références dans [2]) indique que les différences intrinsèques en arabe sont relativement très faibles. Le rapport moyen, qui est de 3 %, n'atteint pas le seuil différentiel de fréquence fondamentale (≈ 6 %) défini pour les sons de la parole [7]. Les différences intrinsèques entre les voyelles longues accentuées semblent plus significatives, puisque le rapport moyen entre voyelle haute et voyelle basse est de 5,5 % (soit une différence absolue de 10,7 Hz).

Le fait que les voyelles longues en arabe semblent avoir des différences intrinsèques plus élevées que celle des voyelles brèves est un résultat intéressant. L'étude récente de Fisher-Jørgensen [8] montre qu'il n'existe pas de différence intrinsèque de Fo en allemand entre les voyelles tendues comme [i, u] et les voyelles relâchées [I, U].Mais en arabe,il n' y a pas de grande différence de timbre entre voyelle longue et voyelle brève sauf entre [a] et [a:] où la voyelle longue [a:] a un F1 relativement plus haut que celui de la voyelle brève [a](d'après nos résultats et voir aussi [9]). Donc la quantitévocalique joue un rôle très important sur la magnitude des différences intrinsèques des voyelles en arabe standard.

3.2 L'EFFET DE LA POSITION DANS LA PHRASE

La figure n°2 et le tableau n°3 montrent que la différence intrinsèque de Fo entre [a:] et [i:] varie significativement en fonction de la position dans la phrase:

- en position initiale le rapport moyen est de 8,4 %, soit une différence de 14,7Hz en valeur absolue.
- en position finale le rapport est très réduit: 1,8 %, soit 2,3 Hz en valeur absolue.

position	initiale	finale
[i:]	190,5	125,2
[a:]	175,8	122,9
Δ Fo en Hz	14,7 Hz	2,3 Hz
[i:]/[a:] en %	8,4 %	1,8 %

Tableau n°3: valeurs moyennes de Fo de [a:] et [i:] en fonction de la position dans la phrase (5 répétitions, 2 locuteurs).

Ces résultats vont dans le même sens que ceux de [4] et [5]: les différences intrinsèques sont beaucoup plus faibles en fin de phrase qu'au début de phrase.

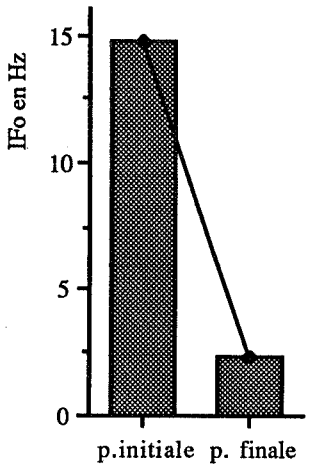


Fig 2: Différence de FoI entre [a:] et [i:] en fonction de la position dans la phrase (2 locuteurs).

3.3 L'EFFET DU FOCUS

Parmi les objectifs de l'étude de l'effet du focus dans deux positions de la phrase (initiale et finale) sur les différences intrinsèques entre [a:] et [i:], l'un est de tester l'hypothèse selon laquelle l'étendue des variations intrinsèques serait liée au registre du locuteur [4, 5]. Cette hypothèse prédit que:

(1) les différences intrinsèques seront plus marquées lorseque le degré du focus est augmenté et Fo est élevée.

(2) les différences intrinsèques s'excercent plus fortement dans la position de la phrase où Fo est plus élevée.

Les résultats sont schématisés dans les figures 3, 4, 5 et 6 où les différences intrinsèques entre [a:] et [i:] sont indiquées pour chaque niveau du focus initial et final. La valeur moyenne de ces différences en fonction des 5 degrés du focus est donnée à côté. Le niveau n° 1 correspond au degré le moins emphatique et le n° 5 le plus emphatique.

Si nous comparons pour chaque locuteur les valeurs intrinsèques en fonction à la fois du focus initial et du focus final, nous constatons que la deuxième prédiction de l'hypothèse selon laquelle l'étendue des différences intrinsèques est liée à l'élévation du Fo est confirmée chez les deux locuteurs. L'hypothèse prédit que les variations intrinsèques seront plus marquées dans le contexte où Fo est très élevée. Ce contexte correspond au focus initial dans la phrase. Nos résultats sont, donc, en accord avec cette prédiction:

- en focus initial le rapport entre la voyelle haute [i:] et la voyelle basse [a:] est de 13,8 % chez My (soit 37,8 Hz en valeur absolue). Chez LQ ce rapport est de 11 % (soit 29,3 Hz en valeur absolue).
- en focus final le rapport est très réduit. Il est de 4,5 % chez MY (soit une différence de 12,2 Hz) et il

est de 2,4 % chez LQ (soit 6,3 Hz en valeur absolue).

Cependant, en comparant les valeurs intrinsèques à chaque degré du focus, nous ne trouvons pas un ordre croissant chez les deux locuteurs. Donc, la prédiction n° 1 de l'hypothèse (les différences intrinsèques seront plus marquées dans la condition où plus le degré du focus est augmenté plus la Fo est élevée) n'est pas confirmée.

Les résultats de cette expérience concordent avec ceux de [6] sur l'anglais. Steele trouve la même tendance: la magnitude des différences intrinsèques obtenue n'est pas en corrélation proportionnelle avec l'augmentation du niveau du focus.

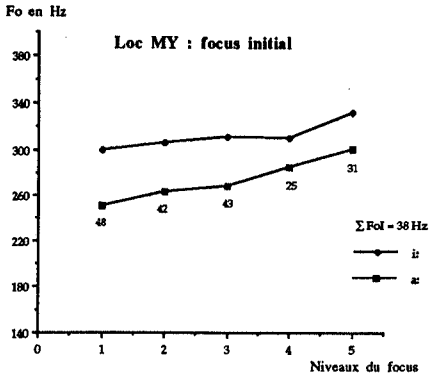
3.4 DISCUSSION GENERALE

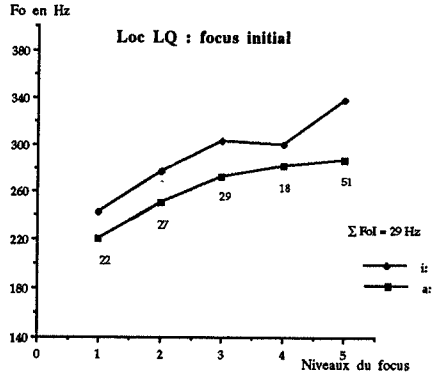
Les résultats de notre travail montrent que le contour général de Fo influe fortement sur les différences intrinsèques entre voyelle fermée et voyelle ouverte. Dans les conditions où Fo est élevée (syllabe accentuée en arabe), en position initiale de phrase, et sous le focus, les différences intrinsèques sont très larges. Ceci suggère qu'une seule explication basée sur les facteurs physiologiques (un surcroît de la tension des cordes vocales dû à l'élévation de la langue: 'hypothèse de l'attraction linguale' [10], ou une activité plus élevée du cricothyroïde [3]) n'est pas suffisante. Plutôt, c'est l'interaction des facteurs physiologiques et aérodynamiques (pression sousglottique [6, 11] qui pourrait expliquer ce phénomène.

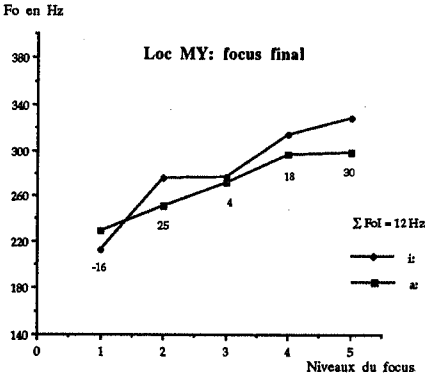
Le fait que qu'il y a une interaction très importante des facteurs prosodiques avec les differences intrinsèques entre les voyelles montre qu'il faut faire une séparation entre les variations de Fo qui sont dues aux caractérestiques segmentales et celles qui sont dues à la prosodie. Une réexamination des effets intrinsèques des voyelles dans différents contextes prosodiques est donc, nécessaire pour une application de l'intonation dans la synthèse et la reconnaissance automatique de la parole. Les algorithmes de correction de Fo doivent prendre en compte cette interaction entre la prosodie et les variations intrinsèques des voyelles.

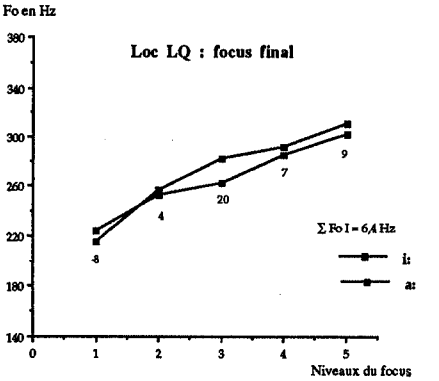
CONCLUSION

Nous avons trouvé que l'accent de mot, la durée vocalique, la position dans la phrase, et le focus jouent un rôle très important sur la magnitude des différences intrinsèques entre les voyelles fermées et les voyelles ouvertes. Ceci montre qu'il faut prendre en compte l'interaction entre le segmental et le suprasegmental dans les variations intrinsèques de Fo. Les résultats mettent aussi en évidence une corrélation générale entre l'élévation du fondamental et les différences intrinsèques de Fo (FoI). Le phénomène de FoI pourrait être dû à l'interaction à la fois des facteurs physiologiques et aérodynamiques.









Figures 3, 4, 5, et 6: Les différences intrinsèques entre [a:] et [i:] en fonction du focus initial et final. Chaque valeur à l'intérieur du graphe représente la moyenne de 3 répétition pour chaque niveau du focus.

REMERCIEMENT

Je tiens à remercier Jacqueline Vaissière pour son aide lors de la rédaction de cette présentation.

BIBLIOGRAPHIE

[1] Silverman, K. (1984) What causes vowels to have intrinsic fundamental frequency?" *Cambridge Pap. Phon. Exp. Ling.* 3: 1-15.

[2] DiCristo, A. (1985) *De la Microprosodie à l'Intonosyntaxe.*Thèse d'état. Université de Provence.

[3] Dyhr, N.-J. (1990) "The activity of the cricothyroid muscle and the intrinsic fundamental frequency in Danish vowels," *Phonetica* 47: 141-154.

[4] Ladd, D. & Silverman, K. (1984) "Vowel intrinsic pitch in connected speech," *Phonetica* 41: 31-40.

[5] Shadle, C.H. (1985) "Intrinsic fundamental frequency of vowels in sentence context," *JASA* 78: 1562-1567.

[6] Steele, S.A. (1986) "Interaction of vowel Fo and prosody," *Phonetica* 43: 92-105.

[7] Rossi, M. & Chafcouloff, M. (1972) "Recherches sur le seuil différentiel de fréquence fondamentale dans la parole," *TIPA* 1: 179-185.

[8] Fisher-Jørgensen, E. (1990) "Fo in tense and lax vowels with special reference to German," *Phonetica* 47: 99-140.

[9] Al Ani, S.H. (1970) *Arabic Phonology.* The Hague. Mouton.

[10] Ohala, J.J. & Eukel, B.W. (1987), "Explaining the intrinsic pitch of vowels," dans Channon, R. & Shockey, K. eds., *In Honor of Ilse Lehiste.* Foris Publications.

[11] Vilkman, E., Raimo, I., Aaltonen, O. (1991) "Is subglottal pressure a contributing factor to the intrinsic Fo phenomenon?" *XIIth ICPhS* 2: 58-61.

19es J.E.P. - Bruxelles, 19 au 22 mai 1992.

LA RECONNAISSANCE DU LOCUTEUR BASEE SUR DES MODELES DE MARKOV CACHES DE PHONEMES

Cl.Vloeberghs (*) et P.Dupont (**)

(*) Ecole Royale Militaire, BRUXELLES
(**) Philips Research Laboratory, BELGIUM

Résumé

On décrit d'abord une méthode de reconnaissance du locuteur, qui est indépendante du texte. Dans cette méthode, chaque phrase correspond à un modèle de Markov caché, que l'on obtient par concaténation des modèles de Markov propres à chaque phonème. Si la suite des états de ce modèle dépend de la phrase, en revanche, les paramètres de ces modèles sont propres au locuteur.

On décrit ensuite le programme qui permet d'évaluer cette méthode et l'on mentionne les paramètres qu'il y a lieu d'ajuster au mieux pour améliorer le taux d'identifications correctes et le taux de rejets corrects des imposteurs.

1. INTRODUCTION

Les méthodes de reconnaissance du locuteur peuvent être classées en deux grandes catégories : celles qui sont dites "dépendantes du texte", et celles qui sont appelées "indépendantes du texte".

Dans le premier cas, le locuteur testé est convié à prononcer la même phrase ou les mêmes mots que lors de la phase d'enrôlement. Les durées, tant de cette - phase d'enrôlement que de la phase de test, sont relativement courtes (de quelques secondes à quelques dizaines de secondes).

Dans le deuxième cas, ces deux phases exigent en général beaucoup plus de temps : comme règle de bonne pratique, on considère qu'une production de minimum une minute de parole (sans silence et sans segments à faible valeur quadratique moyenne) est nécessaire à l'obtention d'un modèle fiable pour chaque locuteur enrôlé [1]. Cependant, les méthodes indépendantes du texte présentent un intérêt certain dans la mesure où les phrases prononcées lors du test peuvent être tout à fait différentes de celles qui ont servi à l'enrôlement.

Un défi dans le développement de méthodes indépendantes du texte consiste à réduire la durée de l'enrôlement ou du test, tout en préservant une valeur convenable du taux d'identification du vrai locuteur.

Cette communication a pour but de :

1°) présenter une méthode de reconnaissance du locuteur indépendante du texte, dont la durée de la phase de test est de courte durée (une phrase de quelques secondes). Cette méthode est basée sur des modèles de Markov cachés, associés aux phonèmes produits par le locuteur;

2°) mettre en évidence les divers paramètres qui doivent être évalués pour utiliser cette méthode de façon optimale.

2. PRESENTATION DE LA METHODE

a. Modèles de Markov cachés

C'est dans le domaine de la reconnaissance automatique de la parole que les modèles de Markov cachés ("HMM: Hidden Markov Models") ont acquis leurs lettres de noblesse [2,3,4]; ils paraissent être

beaucoup moins utilisés pour la reconnaissance du locuteur. Cependant, il convient de mentionner un emploi spécifique de ces modèles pour réaliser une identification de locuteurs, dépendante du texte [5] : chaque locuteur est caractérisé par un modèle de Markov caché à structure circulaire (comprenant 6 états, et 3 transitions possibles par état). Dans cette méthode, le modèle possède une structure fixée une fois pour toutes, et seules les probabilités d'émission et de transition sont propres à chaque locuteur.

b. Principe de la méthode

La méthode présentée ici est également basée sur les modèles de Markov cachés, mais la structure linéaire (de la gauche vers la droite) du modèle n'est pas fixée à priori : elle dépend de la phrase prononcée. La suite des états du modèle correspond à la succession des phonèmes composant la phrase. Chaque phonème est caractérisé par un petit nombre d'états (1, 2 ou maximum 3 par phonème), mais les probabilités d'émission et de transition de ces états sont indépendantes du texte : elles dépendent du locuteur.

Ainsi, chaque locuteur repris dans la base de données est caractérisé par son propre ensemble de modèles de Markov reliés à chaque phonème de la langue utilisée.

Comme chaque phrase peut être considérée comme une concaténation de phonèmes successifs, chaque phrase correspond à un modèle de Markov caché à structure linéaire (de la gauche vers la droite). La suite des états de ce modèle est fixée par le texte, alors que les probabilités de transition et d'émission de ces états dépendent du locuteur.

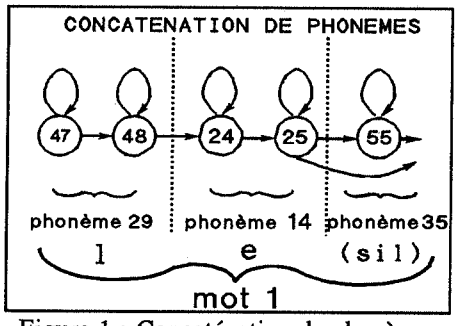


Figure 1 : Concaténation de phonèmes

c. Phase d'enrôlement

Lors de l'enrôlement d'un locuteur, tous les phonèmes rencontrés dans la langue utilisée doivent être prononcés de façon à pouvoir estimer de façon fiable tous les paramètres de leur modèle caché de Markov. Ceci nécessite un ensemble de phrases convenablement choisies, dans lesquelles chaque phonème apparaît plusieurs fois dans des contextes différents.

d. Phase de test

Lors d'un test d'identification, il suffit au locuteur de prononcer une seule phrase, qui peut être tout à fait différente des phrases prononcées lors de l'enrôlement. A cette phrase correspond un modèle de Markov caché, que l'on construit au moyen d'une concaténation des modèles de Markov propres à chacun des phonèmes présents dans la phrase. Ensuite, pour chacun des locuteurs de la base, on calcule la probabilité que la phrase, prononcée par le locuteur à identifier, soit produite par ce modèle de Markov caché, avec les probabilités d'émission et de transition propres à chaque locuteur. C'est le locuteur de la base dont la probabilité est la plus élevée qui est alors "identifié", pour autant que cette probabilité soit supérieure à un seuil approprié, qui permet le rejet des imposteurs.

3. EVALUATION DE LA METHODE

Pour évaluer cette méthode, on a réalisé une base de locuteurs de référence et un programme d'évaluation spécifique.

a. Base de locuteurs

Afin de constituer une base de locuteurs, on a demandé à 21 élèves-officiers francophones de l'Ecole Royale Militaire (8 locutrices et 13 locuteurs) de prononcer 50 phrases, convenablement choisies, au cours de 3 sessions d'enregistrement.

Le texte à prononcer lors de la 1re session était identique pour tous les locuteurs : il s'agissait de 30 phrases en français, phonétiquement équilibrées: chaque phrase a une durée moyenne de 2 à 5 secondes.

La 2e session a été enregistrée directement après la première et correspondait au même texte. Ce sont les 30 phrases de ces deux sessions qui servent à estimer les paramètres des modèles de Markov de chaque locuteur.

La 3e session a été enregistrée 15 jours plus tard. Elle correspondait à la prononciation par chaque locuteur de 2 phrases, répétées 5 fois. Les phrases de cette session différaient de locuteur à locuteur. Elles étaient aussi

tout à fait différentes de celles prononcées lors des deux premières sessions, afin de garantir le caractère d'indépendance du texte de la méthode.

Chaque session d'enregistrement a eu lieu en face d'un poste de travail informatique, produisant un léger bruit acoustique dû au ventilateur, de sorte que le rapport signal-bruit à l'enregistrement variait entre 30 et 40 dB.

b. Traitement préalable des signaux de parole

Le signal de parole correspondant à chaque phrase prononcée est d'abord échantillonné au moyen d'une carte OROS AU-21, à la fréquence de 10 kHz. Les échantillons sont codés de façon uniforme à l'aide de 16 chiffres binaires. Les échantillons successifs d'une phrase sont rassemblés dans un fichier.

Tous ces fichiers sont ensuite traités par des programmes d'usage général, qui permettent:

-d'éliminer automatiquement les deux périodes de silence, au début et à la fin de chaque enregistrement;

-de déterminer, par segment de parole d'une durée de 20 ms, les 16 coefficients cepstraux, et de créer ainsi un nouveau fichier contenant une suite de vecteurs cepstraux par phrase prononcée.

Ces derniers fichiers sont ensuite traités par le logiciel d'évaluation spécifique, dont il est question ci-après.

c. Logiciel d'évaluation

Ce logiciel se compose de deux parties:

-un 1er programme (READ_TREE.C) sert à construire le modèle de Markov (à structure linéaire, de la gauche vers la droite) associé à chacune des phrases utilisées tant pour l'enrôlement que pour le test;

-un 2^{e} programme (RECSPEAKER.C) sert lors de l'enrôlement à estimer les paramètres des états de Markov de chaque locuteur de la base; il sert ensuite pendant la phase de test à obtenir le taux d'identifications correctes des locuteurs de la base et le taux de rejets corrects des imposteurs.

(1) Programme "READ_TREE.C"

(a) Entrées du programme

Ce programme utilise les données reprises dans 3 fichiers.

Le premier fichier, appelé "Phonlist", définit l'ensemble des phonèmes qui sont pris en considération dans la méthode. Dans ce fichier, on attribue à chaque phonème un numéro d'ordre et un certain nombre d'états de Markov (de 1 à 3) .

Le deuxième fichier, appelé "Lexicon_tree", est un dictionnaire dans lequel on reprend tous les mots qui apparaissent dans l'ensemble des phrases utilisées. On attribue à chaque mot un numéro d'ordre, et un certain nombre de prononciations possibles (surtout pour tenir compte des liaisons éventuelles à la fin du mot); ensuite, on indique, par prononciation, le nombre de phonèmes et la succession de ces phonèmes, en les caractérisant par leur numéro d'ordre, tel qu'il a été fixé dans le fichier précédent.

Le troisième fichier, appelé "Sentlist", reprend finalement l'ensemble des phrases utilisées pour l'évaluation. On attribue à chaque phrase un numéro d'ordre, ainsi que la succession des mots de cette phrase, sous la forme des numéros d'ordre définis dans le fichier précédent.

(b) Sorties du programme

Les données de ces trois fichiers permettent au programme "Read_tree.c" de construire, par concaténation des modèles de Markov propres à chaque phonème, un modèle de Markov global pour chacune des phrases utilisées.

En outre, afin de valider les données introduites dans ces trois fichiers, ce programme peut fournir à la demande 4 listes différentes:

-une liste des numéros d'ordre de tous les états des modèles de Markov utilisés dans la méthode;

-la transcription orthographique et l'arbre phonétique de chaque mot (avec toutes les prononciations possibles);

-la transcription orthographique et la description phonétique de chaque phrase, avec le nombre total d'états de Markov par phrase (seule la première

prononciation de chaque mot est imprimée);

-le nombre d'apparitions de chaque phonème dans les phrases d'enrôlement et de test.

(2) Programme "RECSPEAKER.C"

Ce programme comprend deux parties:

-la 1re partie est utilisée, lors de la phase d'enrôlement, pour estimer les paramètres des modèles de Markov;

-la 2e partie est utilisée, lors de la phase de test, pour calculer les statistiques d'identifications correctes des locuteurs de la base et de rejets corrects des imposteurs.

Une fois les paramètres estimés, il est possible de court-circuiter la 1re partie du programme.

(a) Phase d'enrôlement

L'obtention des paramètres des modèles de Markov pour chaque locuteur de la base a lieu au moyen d'une procédure itérative, connue sous le nom d'algorithme de Viterbi.

Pour débuter la procédure, il faut initialiser les paramètres des modèles. On affecte successivement à chaque état des modèles présent dans les différentes phrases traitées, un nombre identique de vecteurs caractéristiques. Ce nombre est égal au nombre de vecteurs de chaque suite, divisé par le nombre d'états du modèle de la phrase correspondante. On peut alors calculer un vecteur caractéristique moyen et une matrice de covariance par état sur l'ensemble des phrases. On obtient ainsi une première estimation des paramètres d'une distribution gaussienne de probabilité d'émission sur chaque état. Les transitions issues des états sont initialement supposées équiprobables.

On peut alors procéder à la première itération. Pour chaque phrase de la première session, on détermine l'alignement optimal entre la suite de vecteurs caractéristiques et les suites admises d'états du modèle de Markov associé à cette phrase. Ces alignements de probabilité maximale définissent une nouvelle affectation des vecteurs caractéristiques de chaque phrase aux différents états. Dès lors, on peut obtenir une nouvelle estimation des paramètres des distributions de probabilité d'émission. Les probabilités de transition sont alors estimées égales aux fréquences relatives de l'apparition de chaque transition le long des alignements optimaux.

Ce processus d'estimation est réitéré jusqu'à obtenir une stabilisation des valeurs des paramètres. Pour notre application, cette stabilisation est atteinte après quelques itérations. Par conséquent, nous avons fixé un nombre maximal d'itérations garantissant la convergence de l'algorithme.

Après avoir utilisé les 30 phrases de la première session pour chacun des locuteurs à identifier, on passe à la détermination du seuil de rejet des imposteurs selon une méthode expliquée plus loin.

On recommence ensuite tout le processus décrit ci-dessus avec les phrases prononcées lors des 2 premières sessions, de manière à affiner l'estimation des paramètres des modèles de Markov.

(b) Phase de test

Pendant cette phase, on utilise les phrases prononcées lors de la 3e session, c'est-à-dire 10 phrases par locuteur. Pour chaque phrase, le programme détermine la probabilité du meilleur alignement entre la suite de vecteurs caractéristiques considérée et les états du modèle de Markov caché relatif à chacun des locuteurs.

C'est le locuteur pour lequel cette probabilité est la plus grande qui est identifié, pour autant qu'il ne soit pas considéré comme un imposteur. Le programme vérifie alors si l'identification est correcte ou si le rejet s'est fait à bon escient; il calcule ainsi, phrase après phrase, le taux d'identifications correctes et le taux de rejets corrects des imposteurs.

d. Détermination du seuil de rejet des imposteurs

La bonne détermination du seuil de rejet des imposteurs est un point important de cette méthode.

Comme toute phrase de test est indépendante des phrases d'enrôlement, et peut varier d'un locuteur à l'autre, il n'est pas possible d'avoir recours à une valeur absolue de ce seuil.

(1) Distance différentielle

On a dès lors préféré utiliser un seuil basé sur une "distance différentielle". Celle-ci est définie à partir des notions suivantes:

-à chaque probabilité finale, déterminée par l'algorithme de Viterbi, on peut faire correspondre une "distance globale" en prenant l'opposé du logarithme de la probabilité (une probabilité maximale correspond ainsi à une distance cumulée minimale);

-lors d'un test, effectué sur une suite de vecteurs caractéristiques, on obtient n distances globales de ce type s'il y a n locuteurs enrôlés dans la base;

-la distance différentielle est définie comme étant la différence entre la plus petite des n distances globales et la moyenne des distances globales d'un groupe de locuteurs choisi parmi les (n-1) locuteurs restants.

Le locuteur dont la distance globale est la plus petite est alors identifié pour autant que la distance différentielle soit supérieure au seuil de rejet.

Le recours à cette distance différentielle repose sur l'hypothèse que la distance globale minimale qui correspond à un locuteur de la base, est plus éloignée de l'ensemble des autres distances globales, qu'une distance globale minimale correspondant à un imposteur.

(2) Détermination du seuil de rejet

Ce seuil de rejet est déterminé pendant la phase d'enrôlement. Après avoir réalisé une première estimation des paramètres avec les phrases de la 1re session, le programme réalise une phase de validation en utilisant quelques phrases de la 2e session. Pour chaque locuteur enrôlé, on calcule une distance différentielle, et c'est la plus petite de ces distances différentielles, sur l'ensemble des locuteurs, qui est choisie comme seuil de rejet pour la phase de test.

4. PARAMETRES ET RESULTATS

Il est évident que les résultats obtenus par cette méthode dépendent d'un grand nombre de paramètres, qu'il y a lieu d'ajuster au mieux.

a. Paramètres à ajuster

Les principaux paramètres à ajuster sont les suivants:

-le nombre d'états dans le modèle de Markov de chaque phonème. On prévoit de faire varier ce nombre de 1 à 3; il est évident que ce nombre doit être adapté à la nature et à la durée du phonème considéré.

-le nombre de transitions issues de chaque état des modèles de Markov. Jusqu'à présent, on n'a envisagé que deux à quatre transitions issues de chaque état.

-le nombre d'itérations lors de la phase d'enrôlement. Après un certain nombre d'itérations sur un ensemble de phrases, on ne détecte plus d'amélioration de la distance globale.

-l'utilisation ou non de composantes différentielles dans les vecteurs caractéristiques de la suite servant tant à l'enrôlement qu'au test.

-la façon de déterminer le groupe de locuteurs qui permet de calculer la distance différentielle définie ci-dessus. Ce groupe peut comprendre soit l'ensemble des (n-1) locuteurs restants, soit un sous-ensemble de ces locuteurs, ceux dont les distances globales sont les plus petites.

b. Résultats

Différents tests ont été effectués sur la base de données décrite plus haut. Les meilleurs résultats obtenus jusqu'à présent, dans un mode indépendant du texte, correspondent à:

92% pour le taux d'identifications correctes des locuteurs de la base,
60% pour le taux de rejets corrects des imposteurs.

Ce résultat a été obtenus avec les paramètres suivants:

- 1 seul état de Markov par phonème, sauf pour les phonèmes (a, er, ou, R, E, o, v, en, s, k, e, t, u, ui, n, m, I, j, l, qui comprennent 2 états);

- 2 transitions issues de chaque état (l'une correspond au bouclage de l'état sur lui-même, et l'autre au passage à l'état suivant, comme indiqué à la figure 1);

-estimation des paramètres à l'aide de 3 itérations

sur les phrases de la session 1, suivies de 3 itérations sur les phrases des sessions 1 et 2.

-pas de composantes différentielles dans le vecteur caractéristique;

-distance différentielle (pour le seuil de rejet) déterminée par rapport à la moyenne des 5 locuteurs ayant la distance globale la plus faible.

D'autres résultats sont mentionnés dans le tableau suivant.

TAUX D'IDENTIFICATIONS CORRECTES (TIC)
TAUX DE REJETS CORRECTS (TRC)

Nb Max états	Nb Max transitions	Nb total itérations	Comp. différ.	Nb locuteurs pour calcul du seuil	TIC (%)	TRC (%)
1	2	6	non	(n-1)	**94**	**0**
2	3	8	non	(n-1)	**88**	**0**
2	2	8	non	(n-1)	**89**	**20**
2	2	6	non	5	**92**	**60**
2	2	10	oui	5	**98**	**16**
2	2	10	oui	17	**31**	**100**

BIBLIOGRAPHIE

[1] R. SCHWARTZ, S. ROUCOS and M. BEROUTI, "The Application of Probability Density Estimation to Text-Independent Speaker Identification," *Proc. ICASSP* PARIS (FRANCE), May 1982, pp. 1649-1652.

[2] H. BOURLARD et al ,"Speaker dependent connected speech recognition via phonemic Markov models", *Proc. ICASSP*, Tampa, pp. 1213-1216, 1985.

[3] K.F. LEE, "*Large Vocabulary Speaker Independent Continuous Speech Recognition: The SPHINX system*" , Ph.D dissertation, Computer Science Department, Carnegie Mellon University, 1988.

[4] P. DUPONT and Y. KAMP, "Guiding speech recognition by a language model", in A. THAYSE (editor),"*From natural language processing to logic for expert systems*", Chap.I, J.WILEY & Sons, Chichester, New York, 1991

[5] Y.C. ZHENG and B.Z. YUAN, "Text-Dependent Speaker Identification Using Circular Hidden Markov Models," *Proc. ICASSP 88* , pp.580-582

19es J.E.P. - Bruxelles, 19 au 22 mai 1992.

PERTINENCE DES TROIS PREMIERS FORMANTS DES VOYELLES ORALES DANS LA CARACTERISATION DU LOCUTEUR

Odile Mella

CRIN-CNRS & INRIA Lorraine
B.P. 239 F54506 Vandœuvre-lès-Nancy CEDEX

Résumé

Le travail rapporté dans cet article se situe dans le cadre de la caractérisation automatique du locuteur. Dans cette optique, nous sommes en train de réaliser une étude sur la pertinence de plusieurs paramètres phonétiques et acoustiques dont le premier volet est constitué par l'étude des trois premiers formants de certaines voyelles orales du français. Il s'agit de déterminer les voyelles orales les mieux adaptées à la reconnaissance automatique du locuteur et, pour chacune d'elles de déterminer les formants ou les combinaisons de formants les plus discriminants. Après avoir décrit l'élaboration et l'étiquetage du corpus dont sont issues ces voyelles, nous développerons la méthode de détermination des trois premiers formants des voyelles. Puis, nous présenterons les indicateurs de pertinence utilisés pour classer les voyelles et les combinaisons formantiques avant de terminer par la présentation de quelques résultats.

INTRODUCTION

La recherche en traitement automatique de la parole a mis en évidence la variabilité du signal de parole en fonction du locuteur. Les études menées dans le cadre de l'identification ou de la vérification du locuteur se classent dans deux catégories selon qu'elles utilisent cette variabilité de façon implicite [ROS, 76], ou qu'au contraire elles essaient de l'extraire du message acoustico-phonétique [CAE, 88]. Notre travail se situe dans cette dernière catégorie. Il a pour but d'étudier la pertinence pour la caractérisation du locuteur d'un certain nombre de paramètres phonologiques, phonétiques et acoustiques spécifiques du français et qui, à terme, pourraient être utilisés dans des systèmes de la première catégorie [GON, 92]. Dans une première étape, nous avons étudié les trois premiers formants de certaines voyelles orales. Nous allons présenter dans cet article, les trois phases de cette étude : l'élaboration et l'étiquetage du corpus nécessaire à l'étude des paramètres, la détermination des trois premiers formants des voyelles orales retenues et le calcul d'indicateurs de pertinence permettant de classer les voyelles orales et les combinaisons formantiques associées. Puis, nous concluons par la présentation de quelques résultats.

ELABORATION ET ETIQUETAGE DU CORPUS

1 Elaboration du corpus

Un ensemble de paramètres fréquentiels, temporels et phonologiques susceptibles d'être pertinents pour la reconnaissance du locuteur a été sélectionné avec l'aide de F. Lonchamp, Professeur à l'Institut de Phonétique de Nancy. Parmi les paramètres fréquentiels figurent les formants des voyelles orales, /ɛ/, /e/, /œ/, /ɔ/, /a/, /i/ et /u/, précédées d'un contexte neutre au sens de la coarticulation linguale, /p/ ou /b/, et suivies d'un contexte postérieur allongeant, /R/. Suite à cette sélection, dix-sept phrases ont été construites de façon à minimiser le nombre de phrases à prononcer par rapport au nombre de paramètres retenus. L'emploi de phrases a été jugé préférable à celui de mots isolés. En effet, la lecture de listes de mots hors contexte devient vite fastidieuse pour un locuteur bénévole et reflète peu sa prononciation naturelle. La table 1 présente les phrases du corpus dans lesquelles les occurrences des triplets /p-voyelle-R/ et /b-voyelle-R/ sont soulignées.

Pour des raisons d'homogénéité de population, dix-sept locuteurs et vingt-et-une locutrices, tous originaires de Lorraine ou y résidant depuis de nombreuses

1 *Guy a péri bêtement du diabète en Italie*
2 *La porte du garage tomba avec lourdeur.*
3 *La partie de belote dura toute la matinée.*
4 *Un bateau à vapeur a quitté le port.*
5 *Le petit gamin traîne un jouet.*
6 *Donne-moi le bocal de cacao !*
7 *En ski, la godille permet d'éviter les tournants.*
8 *Un coq bien dodu pour demain !*
9 *Lequel des bandits guette près du repère ?*
10 *Le trappeur commun redoutait le loup-garou.*
11 *Douze nains conspirent derrière le bosquet.*
12 *Le soldat brisa la baguette de son tambour.*
13 *Goûtez-moi ce cake au beurre !*
14 *Le rire de la gouvernante est revigorant.*
15 *La cousine du nain soupire dans son délire.*
16 *Le départ de la course Strasbourg-Paris aura du retard.*
17 *Notre guide charmant quitte la jolie route danoise.*

Table 1. Les dix-sept phrases du corpus.

années, ont été choisis pour lire quatre répétitions de chaque phrase. Ces répétitions ont été présentées aux locuteurs sous la forme de listes de phrases rangées aléatoirement afin de ne pas toujours reporter sur les mêmes phrases les phénomènes de fin de liste. L'enregistrement, effectué à partir d'un microphone Shure et d'un magnétophone Revox, ne s'est pas déroulé dans une véritable chambre sourde mais dans une pièce isolée et à des heures calmes.

2 Etiquetage du corpus

Une partie du corpus correspondant à dix locuteurs masculins a été numérisée phrase par phrase sur un Masscomp 5600. L'acquisition a été effectuée avec un filtrage passe-bas à 6800 Hz et une conversion A/D sur 12 bits à la fréquence d'échantillonnage de 16 kHz. Puis, les phrases numérisées ont été étiquetées manuellement à l'aide de la version 88 de SNORRI, logiciel interactif d'édition de signal de parole développé au laboratoire [LAP, 88].

La segmentation des phrases, réalisée par l'indication de frontières de segments sur le spectrogramme, a été effectuée suivant un certain nombre de critères. Ces critères ont pour but d'obtenir une segmentation répétitive proche d'une segmentation semi-automatique, de faciliter le calcul des différents paramètres à étudier et de remédier aux hésitations dues au manque d'expérience de l'"étiqueteuse". Ces règles sont nécessaires lorsqu'il y a hésitation entre plusieurs frontières possibles ou lorsqu'il n'existe pas de frontière évidente. Nous donnons ici les critères qui ont été appliqués dans la segmentation de la voyelle du triplet /p-voyelle-R/ :

- souvent, le /R/ est vocalique et donc très difficile à séparer de la voyelle qui le précède. Le positionnement de la frontière s'est surtout fait grâce à l'écoute du segment acoustique, au changement de timbre de la voyelle ;
- lorsque le triplet termine une phrase ou un groupe de phonation, la fin de la voyelle est déterminée par la fin du formant F1 de la voyelle quand ce dernier remonte vers celui du /R/ et que la barre de voisement disparaît ;
- dans quelques cas, par exemple dans la phrase numéro 1, le phonème /R/ est en position intervocalique. L'ensemble /voyelle-R-voyelle/ comporte alors des transitions longues comparativement aux parties stables des phonèmes. Les frontières peuvent être positionnées au début, au milieu ou à la fin de ces transitions. Ces triplets servant aussi à l'analyse spectrale du /R/, il a été décidé d'inclure les transitions dans les voyelles environnantes et de limiter le /R/ à sa partie fricative.

La transcription a été effectuée simultanément à la segmentation et par conséquent à partir des mêmes informations. Elle est de nature à la fois phonologique et acoustique et correspond à ce qui a été entendu et vu, sauf dans les cas ambigus où l'étiquette est celle de l'élément attendu [MEL, 89].

Après cette première phase de construction de la base de données nécessaire à l'étude des paramètres sélectionnés, nous avons étudié la pertinence des trois premiers formants des voyelles orales retenues. Pour cela, nous avons conçu une méthode de détermination automatique de ces formants que nous allons développer dans le paragraphe suivant.

DETERMINATION DES TROIS PREMIERS FORMANTS DES VOYELLES ORALES

Un formant de voyelle orale est déterminé à partir de trois formants intermédiaires calculés à trois emplacements dans la voyelle. Chacun de ces formants intermédiaires est lui-même obtenu grâce à une méthode automatique d'affectation des pôles issus de l'analyse LPC aux formants F1, F2 et F3 de la voyelle orale. Nous allons détailler ces différentes étapes dans les paragraphes suivants, après avoir défini un formant F_i, intermédiaire ou final, comme une structure constituée :

- d'une fréquence formantique, F_i.fr ;

- d'une largeur de bande, F_i.bw ;
- d'un coefficient de défiance, F_i.df ;

Le coefficient de défiance df est composé de trois champs indépendants df_3, df_2 et df_1. Les champs df_2 et df_1 seront mis à jour au cours des différentes phases de calcul des formants finaux alors que df_3 le sera lors du calcul des distances entre locuteurs. Plus le coefficient de défiance est élevé, moins la valeur du formant et la distance calculée à partir de cette valeur sont fiables.

1 Calcul et premiers filtrages des racines LPC

Nous effectuons une analyse LPC fondée sur la méthode d'autocorrélation de Durbin [MAR, 76], qui fournit dix-huit coefficients à partir desquels sont extraites toutes les racines LPC ayant une largeur de bande inférieure ou égale à 1000 Hz. Un premier filtrage des racines LPC permet d'éliminer les pôles ayant des fréquences trop voisines. Si l'écart fréquentiel entre deux racines est inférieur à la valeur médiane du pitch calculé sur la phrase contenant la voyelle étudiée, seule est conservée la racine dont la largeur de bande est la plus petite. Ensuite, un deuxième filtrage des racines est effectué afin de ne conserver que les racines dont la fréquence appartient à un intervalle prédéfini (ou domaine de définition). Ces domaines de définition sont les intervalles fréquentiels dans lesquels sont censés se trouver les trois premiers formants des voyelles retenues, précédées des contextes /p/ et /b/ et suivies du contexte /R/ et prononcées par un locuteur masculin. Ces intervalles ont d'abord été établis sur les conseils d'une phonéticienne du laboratoire puis affinés en fonction des premiers résultats obtenus.

2 Affectation des racines LPC aux formants intermédiaires F1, F2 et F3

Pour chaque voyelle, les racines restantes, triées dans l'ordre croissant de leurs fréquences formantiques, sont affectées séquentiellement aux trois premiers formants F1, F2 et F3 selon le nombre de racines présentes dans le domaine de définition $D(F_i)$ de chaque formant :

- si $D(F_i)$ ne contient aucune racine, le formant intermédiaire F_i est nul ;
- si $D(F_i)$ contient une seule racine candidate, elle est affectée au formant F_i ;
- si $D(F_i)$ contient deux racines susceptibles d'être affectées au formant, un algorithme sélectionne une des deux racines en utilisant l'une des règles suivantes, classées par ordre de priorité :
  - éliminer la racine qui a déjà été affectée à F_{i-1},
  - éliminer la racine qui est la seule candidate pour le formant F_{i+1},
  - éliminer la racine dont la largeur de bande est supérieure au double de la largeur de bande de l'autre racine,
  - éliminer la racine dont la fréquence est la plus éloignée d'une valeur de référence du formant cherché. Nous avons choisi comme valeurs de référence les valeurs médianes des formants F1, F2 et F3 établies par F. Lonchamp dans une étude sur un corpus de triplets /p-Voyelle-R/ [CAL, 89] ;
- si $D(F_i)$ contient trois racines, un algorithme de choix du même type est appliqué.

La largeur de bande du formant, F_i.bw, est donnée par la largeur de bande associée à la racine retenue. De plus, le champ df_1 du coefficient de défiance F_i.df vaut (en Hz)

$$(1) \qquad \begin{cases} 1 & si\ Fi.bw \geq 250 + \dfrac{Fi.fr}{10} \\ 0 & sinon \end{cases}$$

3 Détermination des formants finaux F1, F2 et F3

L'application de la méthode précédemment décrite permet de calculer, pour chacun des formants d'une voyelle orale, trois valeurs intermédiaires à trois endroits situés dans la partie la plus stable de la voyelle, un emplacement central, un emplacement situé à – 8 ms et un autre à + 8 ms. L'emplacement central dépend de la durée de la voyelle, variable selon la position syntaxico-sémantique de la voyelle dans la phrase. Si la voyelle est longue (durée > 160 ms), l'emplacement central est pris à 80 ms du début de la voyelle, sinon il correspond au milieu de la voyelle.

Afin d'obtenir des fréquences formantiques robustes à partir desquelles seront calculées des distances entre locuteurs, l'attribution d'une valeur à une fréquence formantique finale dépend de la proximité des valeurs intermédiaires.

En nous fondant sur les résultats de Monsen et Engebretson [MON, 83] sur les erreurs d'estimation des trois premiers formants des voyelles orales en spectrographie comme en analyse par prédiction linéaire, nous avons choisi, comme écart maximal entre fréquences formantiques, 60 Hz pour F1 et 110 Hz pour F2 et F3. Nous avons utilisé ces écarts pour déterminer la valeur finale du formant et de son coefficient de défiance selon une démarche résumée dans la table 2. On remarque dans cette table que si la fréquence d'un formant n'est pas considérée comme suffisamment fiable, elle est mise à 0.

Une fois les valeurs finales des trois premiers formants des voyelles orales établies, la pertinence pour la caractérisation du locuteur des voyelles orales et de certaines combinaisons de ces formants peut être étudiée. La méthodologie employée pour réaliser cette étude fait l'objet du prochain paragraphe.

ETUDE DE LA PERTINENCE DES VOYELLES ORALES

Le nombre restreint d'échantillons par triplet (quatre par locuteur) ne nous a pas permis d'utiliser une analyse de type analyse discriminante. Aussi avons-nous décidé de limiter les combinaisons formantiques étudiées aux formants (un, deux ou trois formants par voyelle) et aux écarts entre les formants (un, deux ou trois écarts par voyelle).

Lors de l'étude d'une combinaison formantique, pour chaque répétition i de la voyelle étudiée, un locuteur est représenté par un vecteur ayant pour composantes les éléments de cette combinaison. Ainsi, lors du test de la pertinence de la combinaison (F_1, F_2, F_3), le locuteur k est représenté par le vecteur $V_{k_i}(x_1, x_2, x_3) = V_{k_i}(F_1, F_2, F_3)$.

Les méthodes de classement que nous avons mises en œuvre ont recours au calcul d'une distance euclidienne normalisée entre le vecteur associé à l'occurrence i d'une voyelle par le locuteur k et le vecteur associé à l'occurrence j de la voyelle par un locuteur l. A cette distance, donnée par la formule 2 ci-dessous, nous avons adjoint un coefficient de défiance D.df composé de trois champs. Les champs df_1 et df_2 sont respectivement les sommes des champs df_1 et df_2 des coefficients de défiance des composantes du vecteur. Le champ df_3 indique qu'un ou plusieurs termes intervenant dans le calcul de la distance sont nuls en raison de la nullité d'une ou de plusieurs des composantes.

$$(2) \qquad D^2(k_i, l_j) = \frac{1}{N} \sum_{n=1}^{N} \frac{\left(x_n^{k_i} - x_n^{l_j}\right)^2}{a_n^2}$$

avec :

N : nombre de composantes non nulles de chaque vecteur,

a_n : coefficient normalisateur qui, selon les tests effectués, peut prendre les valeurs suivantes :

- le minimum des deux composantes,
- la valeur de référence du formant de rang n ou l'écart entre les valeurs de référence, lorsque les composantes sont des écarts entre formants.
- la largeur du domaine de définition du formant de rang n.

A l'aide de cette distance, nous avons calculé trois indicateurs de pertinence de chacune des voyelles et de chacune des combinaisons formantiques, *Taux* (T), *Score* (S) et *Alpha* (A) :

• *Taux* est le pourcentage de réussite de reconnaissance d'un locuteur parmi dix locuteurs de référence lors de douze expériences de reconnaissance par locuteur. Ces douze expériences ont été obtenues en faisant tourner les différentes répétitions des locuteurs (quatre répétitions de référence et, pour chacune d'elles, trois répétitions "échantillons").

Le locuteur de référence qui minimise la distance au locuteur inconnu est considéré comme reconnu, à condition que l'écart entre cette distance minimale et la deuxième distance la plus faible soit supérieur à un seuil. Sinon, le locuteur reconnu est celui dont la distance a le plus petit coefficient de défiance. Dans les premiers tests, ce seuil a été initialisé à 10% de la distance minimale ;

• *Score* est le cumul, pour tous les locuteurs et pour les douze expériences de reconnaissance, des rangs auxquels les locuteurs ont été reconnus :

proximité des formants intermédiaires	formant final fréquence Fl.fr	formant final champ df2	formant final champ df1
3 fréquences proches	moyenne des 3 fréquences	0	somme des 3 df_1
3 fréquences proches nulles	0	5	0
2 fréquences proches	moyenne des 2 fréquences	2	somme des 2 df_1
2 fréquences proches nulles	0	4	0
2 paires de fréquences proches	moyenne des 3 fréquences ou des 2 meilleures d'entre elles	1	somme des 3 ou des 2 df_1
3 fréquences éloignées	0	3	somme des 3 df_1

Table 2. Détermination du formant final à partir des trois formants intermédiaires.

• *Alpha* est un indicateur de type statistique qui estime de façon indirecte le rapport de la variance intralocuteur à la variance interlocuteur. Il est donné par le rapport de la moyenne des distances intralocuteur à la moyenne des distances interlocuteur.

Nous allons présenter dans le paragraphe suivant les valeurs prises par ces trois indicateurs ainsi que le classement des voyelles qui en résulte.

QUELQUES RESULTATS

1 Etude de la combinaison (F1, F2, F3)

La table 3 fournit les valeurs des trois indicateurs de pertinence ainsi que le classement des voyelles dans le cas où un locuteur est représenté par les trois premiers formants des voyelles orales et lorsque le coefficient normalisateur a_n est donné par le minimum des composantes entrant dans le calcul de la distance.

Ce tableau met en évidence la pertinence de la voyelle /ɛ/ dans la phrase 9, quel que soit l'indicateur. Vient ensuite la voyelle /œ/ dans la phrase 4 qui se trouve en deuxième position ou en troisième position selon l'indicateur. La voyelle /ɔ/ obtient de bons résultats au sens des indicateurs *Score* et *Alpha* mais réalise un score de reconnaissance très moyen. Nous pouvons noter également la très mauvaise place de /u/ quels que soient l'indicateur, la position de la voyelle dans la phrase ou le contexte antérieur.

Du point de vue général, nous pouvons remarquer que les voyelles situées en fin de phrase ou en fin de groupe de phonation, donc plus accentuées, caractérisent mieux le locuteur que celles situées en début d'un mot lexical ou à l'intérieur d'un mot grammatical (ɛ_09 *vs* ɛ_07 et ɔ_04 *vs* ɔ_02). En revanche, les rares comparaisons possibles ne nous permettent pas de conclure sur la pertinence du contexte antérieur /p/ par rapport à /b/.

voyelle	Taux (%)	voyelle	Score	voyelle	Alpha
ɛ_09	62	ɛ_09	0.78	ɛ_09	0.34
œ_04	60	ɔ_04	0.80	ɔ_04	0.35
œ_10	56	œ_04	0.84	œ_04	0.38
i_15	55	œ_13	0.91	œ_13	0.38
œ_13	55	i_15	0.99	i_15	0.40
ɔ_04	55	a_03	1.17	œ_10	0.42
a_03	55	œ_10	1.24	ɔ_02	0.44
ɔ_02	53	ɔ_02	1.27	i_11	0.46
i_11	51	e_01	1.42	a_03	0.47
u_16	50	a_16	1.53	e_01	0.48
ɛ_07	47	i_11	1.55	a_16	0.49
a_16	45	u_16	1.58	u_16	0.53
e_01	43	u_12	1.84	ɛ_07	0.62
u_12	43	ɛ_07	2.27	u_12	0.64
u_08	18	u_08	3.21	u_08	0.72

Table 3. Classement des voyelles pour les trois indicateurs dans un ordre de pertinence décroissante, pour la combinaison (F1, F2, F3).

2 Etude toutes combinaisons confondues

La comparaison des indicateurs *Taux* et *Score* entre toutes les combinaisons de formants étudiées montre que la pertinence des voyelles ɛ_09, œ_04 et œ_13 pour la caractérisation du locuteur est due aux formants F2 et F3. En effet, comme le montre la table 4, la reconnaissance du locuteur fondée sur la combinaison (F2, F3) est plus performante que celle fondée sur (F1, F2, F3). Les résultats obtenus pour chaque formant (un locuteur étant ainsi représenté par un seul formant) mettent en évidence la prépondérance de F3 sur F2 (24% à 33% pour F2 contre 37% à 49% pour F3).

Pour les combinaisons formées à partir des écarts entre formants, les taux de reconnaissance sont moins bons que ceux obtenus lors de l'utilisation des formants alors que les indicateurs *Score* et *Alpha* conservent à peu près les mêmes valeurs. Bien que ɛ_09 occupe toujours une bonne place, la voyelle /œ/ est globalement la plus pertinente quelle que soit sa position dans la phrase. De plus, la table 4 indique un taux de reconnaissance de 70% pour l'occurrence ɛ_04. Notons également la relativement bonne pertinence de la voyelle /e/ qui serait due, d'après l'étude individuelle des écarts, à la différence F2–F1.

combinaison	voyelle	Taux (%)	combinaison	voyelle	Taux (%)
F2 F3	ɛ_09	63	F3-F1 F2-F1 F3-F2	œ_04	72
F2 F3	œ_04	62	F2-F1 F3-F2	œ_04	70
F1 F2 F3	ɛ_09	62	F3-F1 F3-F2	œ_04	68
F2 F3	œ_13	60	F3-F1 F2-F1	œ_13	55
F1 F2 F3	œ_04	60	F2-F1 F3-F2	ɛ_09	54
F1 F2 F3	œ_13	56	F3-F1	u_12	53

Table 4. Les meilleures voyelles au sens de l'indicateur *Taux* pour les combinaisons de formants et pour les combinaisons d'écarts entre formants.

3 CONCLUSION

Nous sommes en train d'effectuer une étude sur la pertinence de certains paramètres phonétiques et acoustiques pour la caractérisation du locuteur. Nous avons commencé par l'étude des formants des voyelles /ɛ/, /e/, /œ/, /ɔ/, /a/, /i/ et /u/, précédées d'un contexte neutre au sens de la coarticulation linguale, /p/ ou /b/ et suivies d'un contexte postérieur allongeant, /R/. Pour cela, nous avons développé une méthode automatique de détermination de formants fiables. Nous avons présenté quelques résultats mettant en évidence la pertinence de certaines voyelles et de certaines combinaisons de formants. Notre objectif est de vérifier ces résultats en faisant varier les conditions d'expérimentation (normalisation, coefficient de défiance, ...), puis de compléter l'étude en remplaçant les formants mis à 0 pour cause de non-fiabilité par des valeurs déterminées à partir des spectrogrammes.

BIBLIOGRAPHIE

[CAE, 88] G. Caelen et G. Pérennou, "Phonetico-phonological, Prosodic and Frequential Analyses in a Both Global and Local Approach to Speaker Identification and Verification", Bull. L.C.P., vol. 2, pp 425–455, 1988.

[CAL, 89] Calliope, "La Parole et son Traitement Automatique", Collection CNET-ENST, Masson, pp. 79–120, Paris 1989.

[GON, 92] Y. Gong and J.P. Haton. "Non-linear vectorial interpolation for speaker recognition", Proceedings IEEE–International Conference on ASSP, San Francisco, USA, March 1992.

[LAP, 88] Y. Laprie, "SNORRI : un système d'étude interactif de la parole", Actes des 17e journées d'Etude sur la Parole, pp. 71–76, Nancy, septembre 1988.

[MAR, 76] J.D. Markel and A.H. Gray Jr., "Linear Prediction of Speech", Springer Verlag, New-York, USA, 1976.

[MEL, 89] O. Mella et M.C. Haton, "Méthodologie d'étude de la pertinence de paramètres phonétiques et acoustiques pour la reconnaissance du locuteur", Séminaire sur la variabilité et la spécificité des locuteurs, Marseille Luminy, juin 1989.

[MON, 83] R.B. Monsen and A.M. Engebretson, "The accuracy of formant frequency measurements: a comparison of spectrographic analysis and linear prediction", Journal of Speech and Hearing Research, 26(1), pp. 89–96, 1983.

[ROS, 76] A. Rosenberg, "Automatic Speaker Verification: A Review", Proc. IEEE, vol. 64, pp. 475–487, April 1976.

19es J.E.P. - Bruxelles, 19 au 22 mai 1992.

ÉTUDE DE LA VARIABILITÉ SPECTRALE POUR LA CARACTÉRISATION DU LOCUTEUR

BONASTRE JEAN-FRANÇOIS, MÉLONI HENRI

LABORATOIRE D'INFORMATIQUE UNIVERSITÉ D'AVIGNON

Résumé

Les systèmes de reconnaissance du locuteur basés sur l'étude des informations spectrales à court terme obtiennent des résultats satisfaisants, mais cette approche met en évidence le problème des différents types de variabilité dans la réalisation des unités phonétiques par un locuteur donné. Pour tenter de diminuer l'influence de ces phénomènes sur les résultats des systèmes d'identification du locuteur, nous avons étudié les limites de la variation intra et inter-locuteurs de chaque phonème selon son contexte dans le but de caractériser au mieux un individu à l'aide d'un nombre minimum de références.

Pour cela, à partir d'un corpus important extrait de la BDSON, nous établissons des statistiques sur les variabilités intra et inter-locuteurs au moyen de plusieurs algorithmes de comparaison de formes spectrales. Avec ces résultats, nous définissons un jeu minimal de paramètres et de références spectrales utilisant de manière optimale les informations caractéristiques du locuteur contenues dans les spectres à court terme.

1 Introduction

L'information permettant de caractériser un individu à partir d'un message vocal se présente sous diverses formes. Les systèmes de reconnaissance automatique du locuteur utilisent essentiellement trois catégories d'informations : les informations phonétiques (réalisations spécifiques des sons), prosodiques (intonation, accentuation, distribution des pauses, etc.) et phonologiques (oppositions phonémiques, liaisons, élisions, réalisation acoustiques du schwa, etc.).

Les connaissances phonologiques spécifiques d'un individu sont difficiles à mettre en œuvre dans un système automatique de reconnaissance du locuteur. Ces informations présentent un caractère important de variabilité, elles résistent à une formalisation utilisable en pratique et nécessitent de nombreux énoncés aussi bien pour extraire les particularités pertinentes d'un individu que pour effectuer son identification.

Pour ce qui concerne les informations prosodiques spécifiques d'un locuteur, les travaux essentiels ont porté sur l'étude des variations de la fréquence fondamentale. Les systèmes qui utilisent des paramètres discriminants indépendants des énoncés (valeurs moyennes, limites et vitesses des variations de F0, etc.) obtiennent de très bons résultats (Rosenberg, 1976). Néanmoins, ces paramètres nécessitent un corpus important constitué d'énoncés particuliers aussi bien pour l'apprentissage que pour la reconnaissance. Malgré l'importante quantité d'informations spécifiques du locuteur contenue dans les paramètres fondés sur une analyse dynamique de la fréquence fondamentale (comparaison de contours mélodiques) (Atal, 1976 ; Sambur, 1975 ; Kraayeveld, 1991), les systèmes utilisant ce type de connaissances sont beaucoup plus délicats à mettre en œuvre. Ces techniques comportent les mêmes défauts que les précédentes et sont très dépendants de la structure syntaxique, du contenu de l'énoncé et de la situation de dialogue. Rosenberg et Lummis (Rosenberg, 1976 ; Lummis, 1972) ont mis en évidence la fragilité des systèmes de type prosodique face à des imposteurs de haut niveau, tel des imitateurs professionnels, ainsi que leur sensibilité aux variations physiologiques du locuteur. En outre, certains travaux portant sur l'importance de l'accent régional dans le processus d'identification (Nolan, 1989, 1990) montrent que les caractéristiques communes au groupe des locuteurs prennent le pas sur les spécificités prosodiques de l'individu.

La recherche de caractéristiques spectrales spécifiques du locuteur a débuté par l'utilisation des spectres moyens à long terme (Pruzansky, 1963). Ces techniques utilisaient une référence par locuteur constituée d'un spectre moyenné, calculé sur un corpus d'apprentissage important. Cette méthode est indépendante de l'énoncé – à condition qu'il comporte suffisamment de phonèmes pour une durée de message de l'ordre de plusieurs

dizaines de secondes – mais obtient des résultats moyens qui réduisent son utilisation à celle de filtre sélectionnant un sous ensemble de locuteurs.

La quantification vectorielle a ouvert une nouvelle voie aux systèmes indépendants de l'énoncé, en exploitant l'information spécifique du locuteur véhiculée dans les spectres à court terme. Durant la phase d'apprentissage, on construit un codebook contenant les prototypes des divers sons pour chaque locuteur (en général seules les voyelles sont prises en compte). Lors de la reconnaissance, les spectres du message sont comparés aux prototypes de chaque codebook, les distances minimales par locuteur et par spectre étant cumulées pour effectuer la reconnaissance. La variabilité spectrale intra-locuteur est prise en compte par l'augmentation de la taille des codebook, ce qui permet d'avoir plusieurs prototypes pour un son. Cependant l'absence d'étiquetage des segments ne permet pas de tirer profit du contexte articulatoire d'une référence pour absorber les phénomènes de coarticulation. Cette technique permet la réalisation de systèmes performants avec de petits ensembles de locuteurs (Soong, 1985, 1988) mais, malgré diverses améliorations (Higgins, 1986), elle impose un important apprentissage et une durée de message de quelques dizaines de secondes. Une approche analytique permet une étude spectrale explicite des divers sons éventuellement adaptée au contexte.

Nous avons développé un premier système analytique d'identification du locuteur relativement indépendant de l'énoncé (Bonastre, 1991) qui obtient des résultats intéressants (pas d'erreur pour 20 locuteurs et 6 à 8 secondes de discours, l'apprentissage automatique nécessite 4 secondes de signal). Des contraintes importantes ont été introduites pour faciliter la localisation et l'identification automatique des phonèmes : énoncés connus par le système, utilisation des seuls phonèmes localisables sans erreur, choix des phonèmes dans un contexte peu déformant. Le système mémorise une référence par locuteur et par phonème et, durant la reconnaissance, seuls les phonèmes proposés dans des contextes limitant la variabilité intra-locuteur sont utilisés. Les distances spectrales sont optimisées en fonction du phonème (Méloni, 1991).

Pour améliorer les résultats et en particulier réduire la durée de discours, il est nécessaire d'utiliser tous les phonèmes quel que soit leur contexte. Nous avons étudié les limites de la variation intra et inter-locuteurs de différents paramètres spectraux pour chaque phonème selon son contexte phonémique. A partir de ces données, nous proposons un compromis entre la durée de l'apprentissage, la taille de la base des références et le nombre de phonèmes énoncés autorisant l'identification d'un locuteur.

2 - Méthodologie de mesure de la variabilité des références spectrales

Pour étudier la variabilité inter et intra-locuteur des spectres associés aux phonèmes, nous avons utilisé un corpus important extrait de la base BDSON, composé d'un énoncé étiqueté ("la bise et le soleil") d'environ 50 secondes prononcé par 22 locuteurs. Les spectres ont été calculés de trois manières différentes : à l'aide d'une transformée de fourier rapide, par une prédiction linéaire à 14 coefficients et à 21 coefficients. Les paramètres, calculés toutes les 10 ms, sont représenté soit par 24 valeurs réparties suivant une échelle de Mel, soit par 128 valeurs réparties linéairement.

2.1 - Mesures de la ressemblance spectrale

Pour le calcul de distances à des vecteurs de référence représentant des unités phonétiques, de nombreuses techniques ont été utilisées dans de multiples contextes. Le choix d'une méthode optimale dépend de nombreux facteurs tels que la qualité du signal, le type de représentation paramétrique, le nombre de vecteurs employés dans le calcul, la prise en compte d'informations concernant l'unité phonétique à identifier et/ou son contexte, etc.

Les distances doivent mettre en valeur les informations les plus pertinentes dans des contextes divers. Compte tenu des systèmes de représentation paramétrique que nous avons choisis, notre travail a consisté à résoudre les problèmes suivants :

- ajustement des niveaux d'énergie spectrale,
- prise en compte optimale des variations de position et d'amplitude des maxima spectraux (formants),
- intégration d'informations contextuelles disponibles dans le signal.

Une manière simple de mesurer la distance entre deux spectres est de calculer la somme de la valeur absolue des écarts entre chacun de leurs canaux. Exemple :

$$d_t = \frac{1}{N} \sum_{i=1}^{N} | Cr_i - Cs_i |$$

où N correspond au nombre de canaux d'un spectre, Cr_i au canal i de la référence et Cs_i au canal i du signal pour la trame t. Cette technique a plusieurs inconvénients dont les deux principaux sont le masquage de particularités spectrales significatives et la difficulté d'un ajustement optimal des énergies.

Une alternative à ces difficultés consiste à effectuer une mesure différentielle (comparaison de la dérivée des spectres) qui met en évidence les mouvements des formants. Dans ce cas l'ajustement des niveaux d'énergie est automatique. Une des distances différentielles utilisées correspond à la formule :

$$d_t = \frac{1}{N-k} \sum_{i=k+1}^{N} \left| (Cr_i - Cr_{i-k}) - (Cs_i - Cs_{i-k}) \right|$$

La valeur de k (nombre de canaux de décalage) est ici égale à 1. Ce type de distance est particulièrement intéressant pour les spectres dont les formants sont nettement marqués (la plupart des voyelles à l'exception des nasales et de /u/). Pour augmenter la dynamique des résultats, le calcul est effectué sur quelques canaux qui maximisent l'écart.

Nous avons aussi utilisé une variante de cette distance différentielle (différentielle-évolution) qui ne tient compte que du signe de la dérivée des spectres :

$$d_t = \text{card}\left(\left\{ i \mid 0<i\leq N,\ \frac{Cr_i - Cr_{i-1}}{Cs_i - Cs_{i-1}} < 0 \right\}\right)$$

Un autre compromis consiste à faire correspondre au mieux les points d'inflexion des spectres ; pour chaque extremum de la référence on mesure la distance de ce point de l'espace spectral au plus proche extremum de même type de la trame de signal. Les écarts sont ensuite cumulés pour obtenir l'indice global de ressemblance (l'énergie globale des deux spectres étant préalablement normalisée).

Plusieurs références par phonème sont mémorisées et différenciées par le contexte phonémique (gauche et droit) dans lequel elles ont été sélectionnées. Le contexte phonémique est utilisé de différentes manières en ne prenant en compte éventuellement qu'une partie de sa caractérisation en terme de traits articulatoires.

2.2 - Situations prises en compte dans les tests

Afin de quantifier la variabilité intra-locuteur des spectres d'un phonème, nous mesurons pour un spectre dans un contexte donné, la distance à la référence correspondante — sélectionnée dans ce même contexte — du locuteur concerné (écarts intra-locuteur). De la même manière, pour évaluer la variabilité inter-individuelle, nous calculons cette même distance aux références des autres locuteurs de la base. Ce travail est effectué pour chaque spectre caractérisant un segment phonétique de l'énoncé.

A partir de ces données, nous avons calculé le pourcentage d'identifications réussies à l'aide d'un seul phonème, les valeurs moyennes et l'écart à la moyenne intra et inter-locuteurs de chacun des paramètres. Nous avons alors calculé le nombre de phonèmes nécessaire à l'identification d'un locuteur pour des coefficients de sécurité de 95,5% et de 99,7%. Les valeurs obtenues définissent le cas le plus défavorable : utilisation de plusieures occurrences d'un même phonème (redondance des informations), pour chaque occurrence on considère que l'imposteur le plus proche de l'individu reste le même, etc.

Remarques :

- durant les tests de reconnaissance, une identification est dite réussie si l'indice du locuteur concerné est strictement inférieur à celui des autres locuteurs de la base,
- il n'a pas été possible, pour conserver un nombre de locuteurs suffisant, de séparer les corpus d'apprentissage et de reconnaissance. Le premier spectre rencontré dans un contexte donné est stocké comme référence, les suivants sont mémorisés dans la base de reconnaissance,

3 - Résultats

La figure 1 montre la relation entre le nombre de références et la prise en compte du contexte phonémique. Dans notre corpus, nous avons rencontré 275 triplets phonémiques sur les 42875 théoriques (il faut que chaque triplet soit présent au moins deux fois pour chaque locuteur : une référence et au minimum un test de reconnaissance). Le nombre de spectres mémorisés dans la base de test (spectres utilisés pour la reconnaissance) est de 7515 lorsqu'on n'utilise pas le contexte et de 2202 sinon. La base de référence a une taille de 5016 octets par locuteur avec un prototype par phonème et 41800 octets par locuteur si on mémorise une référence par contexte.

	1 ref / contexte	1 ref / phonème
nb ref max / loc	42875	35
nb ref / loc	275	33
nb test inter-loc	43797	157752
nb test intra-loc	2202	7515

Figure 1 : nombre de références utilisées, nombre maximum de références possibles et nombre de mesures intra et inter-locuteurs effectuées pour chaque paramètre.

La figure 2 montre, pour chaque phonème, le pourcentage d'identification obtenu avec une seule occurrence de ce phonème, le rapport entre l'écart à la moyenne intra-locuteur et inter-locuteurs du paramètre utilisé (combinaison des distances présentée en 2.1) et le nombre d'occurrences nécessaires pour obtenir un facteur de sécurité de 95,5%. Ces résultats sont présentés suivant deux possibilités de prise en compte du contexte : une référence par phonème suivant son contexte phonémique droit et gauche, une référence par phonème (pas de prise en compte du contexte). Seuls les phonèmes présents en nombre suffisant dans le corpus ont été retenus.

Il est nécessaire de combiner les différents types de distances pour obtenir un paramètre fiable dans toutes les situations et possédant une bonne dynamique. La distance différentielle obtient de meilleurs résultats que la distance simple sur 24 canaux (la position en canaux des formants est relativement stable sur l'échelle de Mel)

mais, avec une échelle linéaire sur 128 canaux, ses performances diminuent, la variabilité intra-locuteur augmentant beaucoup (la position des formants n'est stable qu'a quelques canaux près).

	1 ref / contexte			1 ref /phonème		
pho	id%	σ l/i	95%	id%	σ l/i	95%
f	42%	46%	7	27%	53%	10
s	57%	30%	4	35%	61%	10
v	50%	63%	4	16%	108%	57
z	64%	106%	7	27%	111%	27
p	29%	99%	18	14%	113%	115
t	41%	82%	11	15%	110%	62
k	33%	119%	26	13%	95%	78
b	58%	64%	5	36%	89%	12
d	53%	68%	6	12%	99%	46
m	62%	81%	5	39%	117%	19
l	34%	92%	15	14%	114%	257
r	46%	79%	13	13%	113%	1011
j	84%	71%	2	23%	135%	315
a	53%	71%	6	14%	90%	51
o	50%	102%	8	15%	114%	53
ɛ	60%	66%	3	20%	109%	47
o	70%	95%	11	14%	106%	121
e	51%	68%	6	13%	110%	114
ə	29%	63%	14	18%	114%	55
i	48%	97%	6	23%	111%	40
y	43%	41%	6	10%	69%	43
u	48%	60%	9	23%	107%	58
ã	64%	52%	5	22%	82%	26
õ	61%	86%	4	16%	103%	64
GEN	49%	67%	10	20%	88%	60

Figure 2 : pourcentage d'identification (avec une occurrence), rapport de l'écart type intra-locuteur/inter-locuteurs, nombre de phonèmes nécessaires pour une identification avec un coefficient de sécurité de 95,5%

4 - Conclusion

L'amélioration des performances est très nette lorsqu'on tient compte du contexte phonémique des références, le nombre de phonèmes nécessaires à l'identification d'un locuteur est divisé par un facteur 5 et le taux d'identification avec une seule occurrence d'un phonème est multiplié par 2,5.

Les fricatives sourdes semblent peu perturbées par leur contexte, l'amélioration obtenue ne justifie pas dans ce cas l'accroissement du nombre de références. Les liquides /l/ et /r/ restent, malgré une nette amélioration, trop variables pour une utilisation pratique efficace. Les autres phonèmes présentent un ratio informations spécifiques/variabilité intra-locuteur intéressant même si certains scores obtenus (2 occurrences d'un /j/ pour un facteur de 95,5%, 5 occurrences pour 99,7%) doivent être tempérés par le faible nombre de tests effectués pour certains phonèmes (67 tests intra-locuteur pour le /j/). Pour remédier à ce problème il est nécessaire d'accroitre significativement la taille du corpus et de séparer les corpus d'apprentissage et de reconnaissance.

En pratique, il faut moins de 10 unités phonétiques pour classer les locuteurs d'une base de référence (pour une vingtaine de locuteurs) avec un taux de réussite supérieur à 99,9% lorsqu'on prend en compte le contexte phonémique.

A taille de base de référence égale, un système connaissant un phonème et son contexte possède plusieurs avantages sur des techniques comme la QV. Les résultats ne sont pas perturbés par le nombre de références mémorisées (une seule comparaison par segment phonétique et par locuteur quel que soit le nombre de références alors qu'il y aura autant de comparaisons que de références stockées pour la QV). Les distances spectrales peuvent mettre à profit la connaissance du phonème à traiter et de son contexte.

Nous développons actuellement un système basé sur ces techniques et utilisant un module de localisation automatique des segments phonétiques. Une version préliminaire, testée sur le corpus PEQ de la BDSON (10 locuteurs ayant énoncé 100 phrases) obtient des scores de l'ordre de 39% d'identification réussie à l'aide d'un phonème.

BIBLIOGRAPHIE

B. S. Atal (1976), "Automatic Recognition of Speakers from their Voices" ; *Proc. IEEE*, Vol. 64, n.4, pp. 460-475

J.F. Bonastre, H. Méloni, P. Langlais (1991), "Analytical strategy for speaker identification" ; *Proc. 2nd European Conference on Speech Communication and Technology*, 24-26 septembre, Genova, Italy.

Calliope (1989), *La parole et son traitement automatique*, Masson, pp. 512-543.

K. Choukri, G. Chollet, Y. Grenier (1986), "Spectral transformations through Canonical Analysis for speaker adaptation in ASR" ; *Proc. IEEE ICASSP-86*, Vol. 4, p. 2659.

G. Caelen-Haumont, G. Perennou (1982), "Identification et Vérification du locuteur, timbre de la voix" ; *Rapport DRET*, Vol 1 et 2.

G. Caelen-Haumont, G. Perennou (1988), "A Synopsis of Speaker-Recognition Work Done in France" ; *Bull. L.C.P.* Vol 2, pp. 425-455.

G. R. Doddington (1980), "Voice Identification for Entry Control" ; *Symposium on Voice Interactive Systems*, Dallas.

A. L. Higgins (1986), "Speaker Recognition with Template Matching" ; *Speech Tech'86 Conference*, pp. 273-276, New York, avril.

J. Kraayeveld, A.C.M. Rietveld, V. J. van Heuven (1991), "Speaker characterisation in dutch using prosodic parameters" ; *Proc. 2nd European Conference on Speech Communication and Technology*, 24-26 septembre, Genova, Italy.

R. C. Lummis, A. E. Rosenberg (1972), "Test of an Automatic Speaker Verification Method with Intensively Trained Professional Mimics" ; *JASA*, Vol. 51, n.l.

O. Mella, M. C. Haton (1990), "Méthodologie d'étude de la pertinence de paramètres phonétiques et acoustiques pour la reconnaissance du locuteur" ; *Actes du séminaire "Variabilité et spécificité du locuteur : Etudes et applications"*, Marseille Luminy juin 89, pp. 196-199

H. Méloni, P. Gilles (1991a), "Décodage acoustico-phonétique ascendant" ; *Revue Traitement du Signal*, Vol. 8, n° 2, pp. 107-114.

H. Méloni, P. Gilles (1991b), "Représentation de connaissances indépendantes du locuteur pour la reconnaissance de mots acoustiquement proches" ; *XIIème Congrès International des Sciences Phonétiques*, 19-24 Août 1991, Aix-en-Provence.

F. Nolan (1983), *"The Phonetics bases of Speaker Recognition"*, Cambridge University Press.

F. Nolan (1990), "The limitations of auditory-phonetic speaker identification" ; In: H. Kniffka (ed.), *Texte zu Theorie und Praxis forensischer linguistik*, Tubingen, Niemeyer.

W.J. Barry, C.E. Hoequist, F. Nolan (1989), "An approach to the problem of regional accent in automatic speech recognition" ; *Computer Speech and Language 3*, pp. 355-366.

A. E. Rosenberg, M. R. Sambur (1975), "New Techniques for Automatic Speaker-Verification" ; *IEEE Trans. Acoust., Speech and Signal Processing*, VOL. ASSP-23, pp. 169-176.

A. E. Rosenberg (1976), "Automatic Speaker Verification, a Review" ; *Proc IEEE*, 64, 4, pp. 475-487.

A. E. Rosenberg, K. L. Shipley (1981), "Speaker Identification and Verification combined with Speaker Independent Word Recognition" ; *IEEE ICASSP*, ATLANTA, GA, USA, VOL. 1, pp. 184-187.

M. R. Sambur (1975), "Selection of Acoustic Features for Speaker Identification" ; *IEEE Trans. Acoustics, Speech, Signal Processing*, ASSP-23, pp. 176-182.

F. K. Soong, A. E. Rosenberg (1988), "On the use of Instantaneous and Transitional Spectral Information in Speaker Recognition" ; *IEEE Trans. on Acc., Speech and Signal Proc.*, vol 36, No 6, pp. 871-879.

19es J.E.P. - Bruxelles, 19 au 22 mai 1992.

ANALYSE DE LA VARIABILITE DU SPECTRE A LONG TERME REFLEXIONS METHODOLOGIQUES ET ETUDES DE CAS

A. LANDERCY (1), B. HARMEGNIES (1), J.P. KÖSTER (2), E. ABSIL (1), N. MARTIN (1)

UNIVERSITE DE MONS-HAINAUT (1)
UNIVERSITE DE TRÊVES (2)

Résumé

L'article développe une réflexion méthodologique sur la quantification de la variabilité du spectre moyen à long terme. Celle-ci se double de la relation de deux expériences ponctuelles. La première atteste de l'effet de la langue parlée sur le spectre moyen à long terme dans le chef de sujets bilingues luxembourgeois/français. La seconde investigue les répercussions sur la qualité vocale des variations émotionnelles du locuteur: les sujets, enregistrés avant et après une séance de relaxation présentent une variabilité sensible de leurs spectres à long terme. Les deux expériences sont commentées et critiquées en termes d'améliorations à apporter aux recherches futures en la matière.

INTRODUCTION

Depuis plusieurs années, le Service de la Communication Parlée de l'Université de Mons-Hainaut a principalement centré ses activités de recherche sur l'analyse acoustique de la qualité vocale.

Celle-ci a le plus souvent été objectivée par le recours au spectre moyen à long terme (SMLT), considéré comme un indicateur de la qualité vocale dans sa globalité. A l'occasion de ces investigations, le Service a développé divers outils assurant la quantification de la variabilité des spectres recueillis (Harmegnies, 1988, a; Landercy et Harmegnies, 1986; Harmegnies, 1988, b; etc.).

Diverses sources de variation du SMLT ont ainsi pu être prises en considération. Parmi celles-ci, on notera, entre autres, l'individualité du locuteur: variabilité "résiduelle" du spectre à long terme (Harmegnies, 1988, b), le sexe (Harmegnies et Landercy, 1991), le contenu phonémique (Harmegnies et al., 1991), le "setting" (Harmegnies, Esling et Delplancq, 1989), voire encore la langue parlée (Harmegnies et al., 1987; Bruyninckx et al., 1991; Harmegnies et al., 1989) ou l'état émotionnel du locuteur.

Ce sont précisément ces deux dernières sources de variation dont la présente contribution se propose d'illustrer l'étude. Après un bref aperçu méthodologique relatif à nos techniques habituelles d'investigations, nous nous livrerons donc ici à la relation de deux expérimentations: l'une, centrée sur l'effet du bilinguisme français-luxembourgeois; l'autre, focalisée sur les répercussions de la

variation de l'état de détente du locuteur.

2. MÉTHODOLOGIE

2.1. Dispositif expérimental

Comme la plupart de nos investigations menées dans le domaine, les recherches évoquées ici visent chacune à la mise en évidence de l'effet d'une variable indépendante spécifique (la langue, l'état émotionnel) sur la qualité vocale considérée comme variable dépendante. Un soin tout particulier est dès lors apporté à la maîtrise des variables parasites potentielles.

Les conditions d'enregistrement sont dès lors étroitement standardisées: prise de son en chambre anéchoïque, dispositif de captation sonore de qualité élevée et invariante (microphone Neumann U87I, unité de digitalisation PCM SONY), contrôle de la position du sujet, et en particulier de la distance micro-lèvres (par le recours à un repose-front), etc. Afin de minimiser l'éventuelle variabilité résultant des caractéristiques du corpus produit, les sujets sont en outre priés de prononcer des textes invariants, phonétiquement équilibrés.

La plupart des expérimentations entreprises dans ce domaine -et particulièrement celles relatées ici- privilégient en outre une approche intra-sujet. Seule celle-ci garantit en effet l'annulation des variations inter-sujet, particulièrement importantes en la matière.

2.2. Analyses acoustiques

Les spectres à long terme sont réalisés à partir de spectres instantanés à haute résolution, couvrant l'ensemble de la production envisagée et admis à l'entrée d'un algorithme de moyennage. La bande fréquentielle envisagée couvre, la plupart du temps, la zone conversationnelle.

Les dispositifs utilisés dans le cadre de nos recherches sont tantôt l'analyseur Bruel Kjaer 2033, tantôt la station Kay DSP 5500 voire encore, un ensemble de logiciels *ad hoc* réalisés en collaboration avec la Faculté Polytechnique de Mons et tirant parti de processeurs TMS 320 C 30. Le plus souvent, le système d'analyse calcule le SMLT en temps réel, ce dernier étant ensuite transmis à un ordinateur via une carte d'interface (GPIB ou SCSI); la saisie en mémoire de l'ensemble du signal, suivie d'une analyse en différé est en général évitée.

2.3. Procédure de comparaison

Les expériences présentées ici requièrent la comparaison de spectres à long terme entre eux. Celle-ci nécessite évidemment la quantification des différences observées. La littérature offre divers types de procédés quantitatifs. La distance euclidienne et le coefficient de corrélation sont les plus populaires d'entre eux.

Il apparaît cependant (Harmegnies, 1988, b) que ceux-ci présentent des limitations qui oblitèrent leur validité (nécessité de normaliser l'intensité dans le cas de la distance euclidienne; défauts sélectifs de sensibilité dans le cas du coefficient de corrélation).

Une nouvelle technique quantitative a dès lors été mise au point (Harmegnies, 1988, a). Elle consiste en le calcul d'un indice de dissimilarité: l'indice SDDD (Standard Deviation of the Differences Distribution). Ce dernier s'est avéré d'une puissance comparable à celle des mesures classiques; il n'en présente cependant pas les limitations (Harmegnies, 1988, a; Shevchenko et Skopintseva, 1991).

3. EXPÉRIENCE 1.

3.1. Objet

Cette investigation (Martin, 1991) s'inscrit dans le domaine des recherches visant à objectiver les effets éventuels de la langue parlée sur le spectre à long terme. Les langues envisagées sont le luxembourgeois et le français.

3.2. Dispositif

24 sujets ont fait l'objet de cette recherche (15 hommes, 9 femmes). Tous étaient étudiants à l'Université de Trèves (Allemagne). Chacun pratiquait couramment les deux langues envisagées ainsi que l'a révélé un questionnaire socio-linguistique.

Chaque locuteur a réalisé 5 productions d'un texte français équilibré et d'un texte luxembourgeois. Les enregistrements ont été réalisés sur DAT au laboratoire de phonétique de l'Université de Trèves.

Leur analyse fut ultérieurement réalisée au Service de la Communication Parlée au moyen de l'analyseur Bruel Kjaer 2033, celui-ci produisant des spectres à long terme de 400 canaux, présentant une résolution constante de 12.5 Hz dans l'ensemble de la bande d'analyse (DC-5 kHz).

Pour chacun des 24 sujets, 45 comparaisons interspectrales furent réalisées: 10 comparaisons intra-langue sur base des productions françaises, 10 comparaisons intra-langue sur base des productions luxembourgeoises, et enfin 25 comparaisons inter-langue impliquant les productions françaises et luxembourgeoises. Chacune de ces comparaisons a conduit au calcul d'une valeur d'indice SDDD.

3.3. Résultats

L'ensemble des valeurs ainsi obtenues est résumé au tableau 1, qui présente, pour chaque sujet et chaque type de comparaison, la moyenne des indices SDDD ainsi calculés. Comme le révèle l'observation de ce tableau, l'indice moyen inter-langue est, pour chacun des locuteurs, supérieur aux valeurs intra-langue correspondantes.

Loc.	F/F	L/L	F/L
1	2.43	2.01	2.87
2	3.04	2.25	3.52
3	2.35	2.71	3.68
4	2.92	2.22	3.97
5	2.96	1.72	3.44
6	2.86	2.04	4.63
7	2.79	1.93	3.69
8	2.68	2.05	3.76
9	2.44	1.99	3.19
10	2.72	2.47	3.40
11	2.77	2.97	3.79
12	2.29	2.11	3.03
13	2.52	1.57	3.99
14	2.20	2.13	3.87
15	1.94	1.62	3.17
16	3.09	2.37	3.42
17	2.57	2.16	3.69
18	2.91	2.09	3.65
19	2.67	2.40	3.43
20	2.70	2.18	3.98
21	3.20	3.00	3.95
22	2.66	2.13	3.47
23	3.09	2.55	3.82
24	2.92	2.12	4.11

Tableau 1: Moyennes par locuteur ("Loc.") des indices SDDD issus des comparaisons intra-langue (français/français: "F/F" et luxembourgeois/luxembourgeois: "L/L") et inter-langue ("F/L").

Il apparaît clairement que la variabilité vocale introduite par le changement de langue est supérieure aux variabilités inter-production observables dans chacune des langues. Autrement dit, les spectres à long terme issus des deux langues prises ici en considération sont plus dissimilaires entre eux que les spectres à long terme issus d'une seule et même

langue (soit le français, soit le luxembourgeois).

Enfin, nous pouvons également constater qu'à deux exceptions près, l'indice de dissimilarité est plus faible en luxembourgeois qu'en français; nous pouvons supposer que ce phénomène relève soit d'un effet de la langue parlée, soit d'un effet d'interaction entre les deux langues.

4. EXPÉRIENCE 2.

4.1. Objet

Cette recherche (Absil, 1991) tente de déterminer si une variation d'état émotionnel (état normal - détente) peut occasionner des variations de la qualité vocale, objectivables par le biais du spectre à long terme.

4.2. Dispositif

Dix-neuf sujets (10 hommes, 9 femmes) ont pris part à l'expérience. Chacun fut enregistré en une seule session dans deux conditions différentes: d'une part, dès son arrivée au laboratoire ("état normal"), et d'autre part, après une séance de relaxation par sophronisation simple ("état détendu"). Dans chacune des deux situations, le sujet réalisa 5 productions d'un corpus français équilibré.

Chaque production ainsi recueillie fut ultérieurement traitée au moyen de l'analyseur Bruel Kjaer 2033, utilisé dans la même configuration que celle présidant à l'expérience précédente. Les spectres obtenus furent également comparés à l'aide de l'indice SDDD.

Trois types de comparaison furent menées à bien: deux types de comparaison intra-condition (normal/normal et détendu/détendu) et un type de comparaison inter-condition (normal/détendu). Chaque type de comparaison intra-condition fut l'occasion, pour chaque locuteur, du calcul de 10 indices SDDD, alors que les comparaisons inter-condition produisirent, pour chaque sujet, 25 indices.

Loc.	N/N	D/D	N/D
1	1.952	2.158	2.682
2	2.059	1.851	2.772
3	2.309	2.235	3.479
4	2.148	2.203	3.101
5	2.596	2.671	3.957
6	2.696	3.538	4.142
7	2.460	2.743	4.360
8	2.103	2.357	4.642
9	2.170	2.236	3.502
10	2.557	2.412	3.645
11	3.155	2.947	4.015
12	2.639	2.497	3.739
13	3.077	3.120	4.652
14	2.535	2.519	3.165
15	3.527	4.087	4.322
16	3.656	4.120	4.219
17	2.821	2.580	3.711
18	3.179	3.282	3.414
19	2.540	2.388	2.826

Tableau 2: Moyennes par locuteur ("Loc.") des indices SDDD issus des comparaisons intra-condition (normal/normal: "N/N" et détendu/détendu: "D/D") et inter-condition (normal/détendu: "N/D").

4.3. Résultats

Les valeurs ainsi obtenues ont été résumées dans le tableau 2, qui présente pour chaque sujet, la moyenne des indices SDDD obtenus à l'occasion de chacun des types de comparaison. L'examen du tableau révèle que, pour chaque sujet, la valeur de SDDD relative aux comparaisons inter-condition est supérieure aux valeurs intra-condition correspondantes. On peut donc en conclure qu'il existe plus de différences entre les SMLT produits dans deux conditions différentes qu'entre ceux produits dans une situation déterminée. Ces données semblent donc attester d'un effet de la variation d'état émotionnel sur la qualité vocale. Remarquons également que,

en fonction de l'état émotionnel (colonnes 1 et 2 du tableau), la variation de l'indice est fortement dépendante du sujet: 11 sujets sur les 19 présentent en effet une augmentation en situation de détente alors que 8 sujets présentent une diminution.

5. CONCLUSIONS

Les effets de deux variables indépendantes (langue parlée et état émotionnel) sur le SMLT, considéré comme variable dépendante mesurant la qualité vocale, ont pu être mis en évidence.

En ce qui concerne la langue, après élimination des différences inter-individuelles, les résultats nous permettent d'affirmer que la variabilité inter-langue est toujours supérieure à l'intra-langue. Il est à noter également que dans 22 cas sur 24, la variabilité du SMLT apparaît moins importante dans une langue (le luxembourgeois) que dans l'autre. Cette constatation mériterait sans doute des développements ultérieurs; ils pourraient, par exemple, chercher à déterminer si l'on peut trouver là un éventuel effet propre à la langue elle-même, ou plutôt à la maîtrise de celle-ci par le sujet, voire encore à l'influence, dans son chef, d'une langue sur l'autre. Il faut également noter que la contamination d'une troisième langue (l'allemand, en l'occurence) a pu jouer un rôle non contrôlé dans ce cas précis.

L'expérience relative aux variations d'états émotionnels est, quant à elle, intéressante par son originalité: les recherches en la matière ont en effet classiquement eu recours à des tentatives d'*accroissement* plutôt que *réduction* du stress. Les travaux futurs semblent pouvoir être poursuivis dans cette voie. Les importantes différences de réaction de locuteur à locuteur mériteraient également un examen spécifique, si possible doublé d'une objectivation externe de l'effet du traitement sur le sujet.

6. REFERENCES

ABSIL, E., "Contribution à l'étude de l'impact de la détente sur la qualité vocale. Effets d'une sophronisation simple", mémoire FSPP, Université de Mons, 1991.

BRUYNINCKX, M., HARMEGNIES, B., LLISTERI, J., POCH, D., "Effects of language change on voice quality. An experimental study of Catalan-Castilian bilinguals", *Acts of the 12th International Congress of the Phonetic Sciences*, Aix-en-Provence, 1991, 2, 182-186.

HARMEGNIES, B., "SDDD, a new dissimilarity index for the comparison of speech spectra", *Pattern Recognition Letters*, 8, 1988, a, 153-158.

HARMEGNIES, B., "Contribution à la caractérisation acoustique de la qualité voale. Analyses plurielles de Spectres Moyens à Long Terme de parole", dissertation doctorale, FSPP, Université de Mons, 1988, b.

HARMEGNIES, B., BRUYNINCKX, M., LLISTERI, J., POCH, D., "Effects of language change on voice quality. An experimental contribution to the study of the Catalan-Castilian case", *Acts of the European Conference on Speech Communication and Technology*, Paris, 1989, 489-492.

HARMEGNIES, B., BRUYNINCKX, M., LLISTERI, J., POCH, D., "Effects of language change on voice quality in bilingual speakers", *Proceedings of the second European Conference on Speech Communication and Technology*

(Eurospeech '91), Genova, 1991, 1, 165-168.

HARMEGNIES, B., ESLING, J., DELPLANCQ, V., "Quantitative study of the effects of settings changes on the LTAS", *Acts of the European Conference on Speech Communication and Technology*, Paris, 1989, 132-149.

HARMEGNIES, B., LANDERCY, A., "Etude acoustico-statistique de la qualité vocale", *Journal d'acoustique*, 1991, 4, 81-90.

HARMEGNIES, B., LANDERCY, A., BRUYNINCKX, M., "An experiment in inter-languages speaker recognition using the SDDD index", *Proceedings of the Eleventh International Congress of the Phonetic Sciences*, Tallinn (U.R.S.S.), août 1987, 249-253.

LANDERCY, A., HARMEGNIES, B., "Quantification interlocuteur de la variabilité spectrale interlangue", *Proceedings of the 12th International Congress on Acoustics*, Vol.I, A1-2, Toronto, juillet 1986.

MARTIN, N., "Contribution à l'étude de l'impact de la langue sur la qualité vocale: le cas du bilinguisme luxembourgeois-français", mémoire FSPP, Université de Mons, 1991.

SHEVCHENKO, T.I., SKOPINTSEVA, T.S., "Effects of social and regional backgrounds on LTAS in British English", *Proceedings of the second European Conference on Speech Communication and Technology (Eurospeech '91)*, Genova, 1991, 1, 169-172.

19es J.E.P. - Bruxelles, 19 au 22 mai 1992.

ANALYSE DE LA VARIABILITE PHONETIQUE EN PAROLE SPONTANEE REFLEXIONS METHODOLOGIQUES ET ETUDES DE CAS

D. POCH (1), B. HARMEGNIES (2), L. AGUILAR (1), M.J. MACHUCA (1), G. MARTINEZ (1)

UNIVERSITE AUTONOME DE BARCELONE (1)
UNIVERSITE DE MONS-HAINAUT (2)

Résumé

Cet article présente 3 expériences ponctuelles réalisées dans le cadre d'une ligne de recherche spécifique développée par les deux institutions signataires: l'analyse des effets du style de parole dans le cadre des langues romanes. Sont ici abordés les systèmes vocaliques de l'espagnol et du français, ainsi que les consonnes nasales de l'espagnol. Les résultats obtenus suggèrent d'importantes variations des sons envisagés sous l'effet du style du parole. L'accroissement de la variabilité des mesures en parole spontanée apparaît omniprésent, alors que d'autres phénomènes, bien que classiquement reconnus, semblent moins universels.

INTRODUCTION

Longtemps marquée par la conception saussurienne d'une "linguistique de la langue", la phonétique n'a que tardivement, dans son histoire, abordé la problématique de la *variabilité*. Les dernières décennies ont cependant connu un intérêt croissant pour l'étude de diverses sources de variation censées influer sur les réalisations phoniques. L'une d'entre elles, le *style de parole*, n'a pourtant que très récemment fait l'objet d'investigations systématiques. Celles-ci se sont néanmoins, dans leur toute grande majorité, centrées sur des langues extérieures au groupe roman. Les connaissances en la matière sur l'espagnol castilian, le portugais, le catalan, l'italien, voire même le français apparaissent ainsi sporadiques, sinon simplement inexistantes.

Les services représentés dans le cadre de cette contribution ont récemment joint leurs efforts pour aborder en commun cette problématique. Nous présentons ici, à grands traits, les points saillants des méthodologies mises en oeuvre; nous détaillons également, à titre illustratif, trois expériences ponctuelles réalisées en commun. La première s'attache à la dynamique des timbres vocaliques en espagnol castilian; la seconde s'interroge sur l'existence de phénomènes de ce type dans une langue à système vocalique beaucoup plus élaboré: le français; la troisième enfin constitue une première approche de ces mécanismes dans le cas des consonnes.

Les trois expérimentations relatées ici présentent un caractère nettement exploratoire. Elles cherchent plus à mettre en évidence *l'existence* de phénomènes qu'à en établir les lois.

2. METHODOLOGIE

Les travaux rapportés ici s'inscrivent dans la ligne des recherches classiques centrés sur la dynamique des systèmes phonémiques. Au contraire de certains de ceux-ci, tantôt basés sur des lectures de logatomes (Lobacz, 1977), tantôt sur celle de listes de mots (Nord, 1974; Nord, 1986), voire encore de phrases (Lobacz, 1976; den Os, 1985), ils cherchent cependant à opposer les productions dérivées de deux formes très contrastées d'expression: la *parole spontanée*, d'une part, et la *parole de laboratoire*, d'autre part.

Dans ce contexte exploratoire, nous avons préféré une approche mono-locuteur assortie d'une analyse quantitative très approfondie, à une focalisation multi-locuteurs, nécessairement restreinte à un plus petit nombre de mesures par sujet.

Chacun des locuteurs envisagés entretenait des relations suivies soit professionnelles, soit amicales avec son interlocuteur. Dans chaque cas, ce dernier était un locuteur natif de la langue utilisée. Dans un premier temps, le locuteur était prié de se soumettre à un entretien semi-directif, au cours duquel l'expérimentateur, dans un climat de conversation détendue, l'invitait à évoquer divers thèmes proches de ses intérêts personnels (lieu de naissance, petite enfance, service militaire, situation familiale, problèmes professionnels ou scolaires). La durée habituelle de cette session était d'environ une heure. C'est le produit de ces entretiens qui constitue les corpus de *parole spontanée.*

Les corpus de *parole de laboratoire* furent recueillis dans un second temps. Quelques jours après les enregistrements, le locuteur était prié de s'adresser au laboratoire de phonétique afin d'y subir une expérience. La tâche proposée consistait pour lui en la lecture de listes de mots, résultant d'un échantillonnage aléatoire pratiqué dans le corpus de parole spontanée, préalablement transcrit et représenté sous forme informatique.

Pour chacune des expériences, le même matériel fut utilisé lors des deux séances d'enregistrements (Université de Barcelone: microphone Shure 515SB unidyne et enregistreur Revox A77; Université de Mons: microphone Neumann U87I et unité de digitalisation PCM SONY connectée à un magnétoscope de haute qualité).

Les mesures segmentales réalisées dans le cadre de ces investigations se basent sur l'analyse spectrographique ordinaire des sons étudiés (sonagrammes à bande large); elles se doublent, en cas d'incertitude, d'une investigation spectrale (FFT à haute résolution: 10Hz). Les analyseurs utilisés furent tantôt le sonagraphe DSP 5500 KAY (Université de Mons), tantôt le système Mac Speech Lab (Université de Barcelone).

Plusieurs mesures sont effectuées sur chaque segment, à savoir celles de ses valeurs formantiques initiales, médianes et finales ainsi que sa durée. Les résultats organisés en forme de base de données informatisée, furent, dans tous les cas, traités au moyen de procédures statistiques extraites des logiciels SPSS. Dans le présent article, seules sont prises en considération les valeurs *centrales* des formants étudiés.

Dans l'ensemble des investigations du même type que celles présentées ici, nous attachons en outre un soin particulier à la neutralisation de toutes les variables parasites dont l'action pourrait être confondue avec celle qui fait l'objet de nos préoccupations: le *style de parole.* Ainsi, par exemple, certains mots aléatoirement extraits des corpus de parole spontanée ont

dû être écartés du corpus de parole de laboratoire car leur isolation dans les listes en parole de laboratoire risquait de modifier certaines caractéristiques des phonèmes ciblés (e.g., modifications d'accentuation). En outre, afin de limiter au maximum les effets de liste qu'aurait pu entraîner la situation de parole de laboratoire, nous avons recouru à un système de présentation *ad hoc*: les mots à prononcer apparaissent seuls, en succession, sur un écran vidéo placé devant le sujet. Leur rythme d'affichage est contrôlé de manière à éviter l'installation de routines prosodiques régulières dans le chef du locuteur; des pauses lui sont en outre régulièrement suggérées et il décide lui-même du moment de reprise de la tâche.

3. EXPERIENCE 1

3.1. Objet

Les études classiques centrées sur d'autres langues que l'espagnol (Lobacz, 1976; Nord, 1974, 1986; den Os, 1985; etc.), indiquent, pour la plupart, une tendance des timbres vocaliques à se rapprocher de la qualité du schwa. Aucun phénomène de ce type n'ayant été mis en évidence dans le cas de l'espagnol, nous avons, dans un premier temps, tenté d'éprouver l'hypothèse d'une *centralisation* de ses voyelles en parole spontanée. Nous nous sommes, en d'autres termes, demandés si le timbre des voyelles, en parole spontanée, pouvait être considéré comme plus proche de celui de schwa que le timbre des voyelles en parole de laboratoire. Un indice spécifique - l'indice δ (Harmegnies et Poch, 1991)- a ainsi été mis au point. Il repose sur l'estimation des distances euclidiennes entre les réalisations étudiées et schwa:

$$\delta = [(F_1-500)^2+(F_2-1500)^2]^{\frac{1}{2}} - [(f_1-500)^2+(f_2-1500)^2]^{\frac{1}{2}}$$

avec F_1 et F_2 les premier et second formants de la voyelle en parole de laboratoire et f_1 et f_2 les premier et second formants de la voyelle en parole spontanée. L'indice δ présente des valeurs nulles si la position de la réalisation dans l'espace F1/F2 est invariante, et positives si la réalisation en parole spontanée est plus proche de schwa que la réalisation en parole de laboratoire.

Dans le cadre de la présente expérience, 20 réalisations de chaque voyelle de l'espagnol ont été prises en considération, dans chaque style de parole. Le tableau 1 présente des résumés statistiques des distributions de valeurs d'indice δ ainsi obtenues.

voyelle	m	σ	N
/i/	155.6110	92.9369	20
/e/	230.4406	125.0558	20
/a/	15.4158	110.9498	20
/o/	101.3282	162.9051	20
/u/	48.2482	115.9172	20

Tableau 1: moyennes (m), écarts types (σ) et effectifs (N) des indices δ, calculés sur base des enregistremens hispanophones.

Ainsi que le montre le tableau 1, les enregistrements de parole spontanée sont dominés par une tendance sensible à la centralisation des voyelles. Les valeurs de δ sont en effet toutes positives et assez élevées. On notera cependant un renforcement de la tendance pour les voyelles d'avant, par opposition aux voyelles d'arrière, ainsi qu'un minimum de centralisation pour /a/.

D'autres traitements, que nous ne pouvons reproduire ici, indiquent en outre que le phénomène s'accompagne d'un considérable accroissement de la variabilité des valeurs formantiques en parole spontanée.

4. EXPERIENCE 2.

4.1. Objet

Les phénomènes mis en évidence dans le cas de l'espagnol apparaissent au sein d'un système vocalique assez dépouillé.

L'expérience présente cherche à déterminer si une langue à système vocalique plus complexe (en l'occurrence, le français) peut présenter le même type de dynamique sous l'effet des mêmes causes. L'expérience précédente est donc ici répliquée, sur base des enregistrements d'un locuteur francophone (Poch et Harmegnies, 1992).

4.2. Observations

Le même schéma expérimental et les mêmes modalités quantitatives que celles caractérisant les corpus hispanophones ont été utilisés ici. Une vingtaine de réalisations de chacune des voyelles orales du français ont été prises en considération dans chaque type de parole. Les valeurs d'indice ainsi obtenues sont présentées au tableau 2.

Comme on peut l'observer à la lecture de ce tableau, la situation est ici extrêmement différente.

D'une part, la plupart des valeurs sont négatives: on ne constate donc pas, dans le chef des voyelles françaises, la même tendance à la centralisation que présentaient les voyelles espagnoles. Il est cependant clair que les réalisations se modifient sous l'effet du style, les valeurs de δ étant non nulles. D'autre part, les valeurs de l'indice semblent globalement plus faibles en français qu'en espagnol.

voyelle	m	σ	N
/œ/	-27.9840	116.0040	19
/i/	-8.9017	142.5077	20
/e/	-4.7390	131.6799	20
/ɛ/	-3.3248	89.0723	20
/a/	-25.2782	73.2550	20
/ɔ/	-26.4928	126.0987	20
/o/	-7.3369	123.0283	20
/u/	-41.6093	115.1560	20
/y/	10.6634	121.9933	20
/ø/	-40.6530	79.8630	19

Tableau 2: moyennes (m), écarts types (σ) et effectifs (N) des indices δ, calculés sur base des enregistremens francophones.

Ajoutons également que les modifications observées sous l'effet du changement de style de parole s'accompagnent d'un important accroissement des aires de répartition des voyelles prises en considération. Ce phénomène a été confirmé par le biais d'une analyse discriminante. L'ensemble des mesures de parole de laboratoire a ainsi été analysé, les formants étant considérés comme variables discriminantes, et les catégories vocaliques comme variables de groupement. Une fois obtenues les fonctions discriminantes, une tâche de reconnaissance a été simulée; elle consistait à associer chaque observation réelle à une catégorie vocalique. Ce traitement a produit 58.6% de reconnaissance correcte en parole spontanée contre 80.3% de reconnaissance correcte en parole de laboratoire (Poch et Harmegnies, 1992).

5. EXPERIENCE 3

5.1. Objet

Les phénomènes mis en évidence à l'occasion des deux expériences relatées ci-dessus concernent essentiellement les *systèmes vocaliques*. L'étude rapportée ici s'interroge sur l'existence éventuelle de mécanismes du même ordre qui affecteraient les *consonnes*. Trois consonnes nasales de l'espagnol castilian, /m/, /n/, /ɲ/. Dans tous les cas envisagés ici, la consonne nasale est située dans une structure VCV.

5.2. Observations

Nous avons limité, dans le cadre de cette démarche exploratoire, nos investigations aux seuls premier formant et formant de nasalité. Des résumés statistiques ont été établis à partir des valeurs recueillies. Ils sont graphiquement représentés aux figures 1, 2 et 3. Chacune de celles-ci présente en parallèle les données issues de la parole spontanée et

celles dérivées de la parole de laboratoire. L'axe d'ordonnée, comme dans un spectrogramme, est celui des fréquences (en Hz). Pour chacun des deux formants pris en considération, chaque graphique présente sa moyenne (point cerclé), son écart type (segment en traits pleins), et l'étendue de la distribution (traits fins).

Ainsi que le montre la figure 1, la valeur du premier formant de /m/ s'accroît en parole spontanée, alors que celle de son formant de nasalité décroît légèrement. Les variations les plus spectaculaires ne touchent cependant pas les valeurs moyennes, mais plutôt la dispersion caractérisant chacun des deux styles de parole: on remarquera que non seulement les écarts types (tant du premier formant que du formant de nasalité), mais aussi les étendues des distributions des valeurs formantiques sont notablement plus importants en parole spontanée qu'en parole de laboratoire. Le même type de tendance est observable à l'examen de la figure 2, relative à la consonne /n/: les valeurs moyennes du premier formant tendent à s'accroître en parole spontanée, alors que celles du formant de nasalité tendent à décroître dans ce même style de parole. Les dispersions sont de manière générale, nettement plus importantes en parole spontanée qu'en parole de laboratoire. Les tendances centrales des valeurs formantiques caractérisant la consonne /n/ évoluent de la même manière que celles des phonèmes préalablement étudiés (en parole spontanée: accroissement du premier formant, décroissance du formant de nasalité). La variabilité des mesures semble également s'accroître - mais de manière plus discrète - au passage de la parole de laboratoire en parole spontanée.

Il apparaît clairement, à la lecture des observations rapportées ci-dessus, que l'accroissement de la *dispersion* des valeurs formantiques est le phénomène le plus

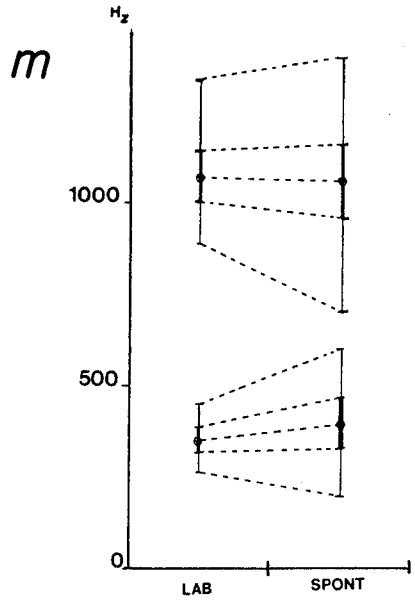


FIG 1

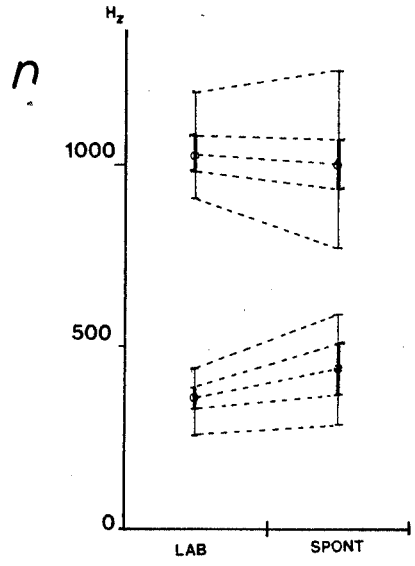


FIG 2

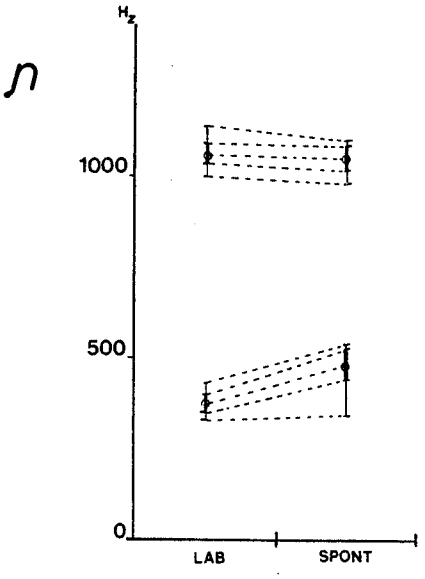


FIG 3

saillant qui se se dégage de l'analyse des corpus de la parole spontanée. Les données fournies au tableau 3 apportent une confirmation plus quantitative à cette constatation. Pour chaque formant et chaque consonne, ce tableau présente le rapport des variances caractérisant les distributions statistiques des fréquences relevées.

	/m/	/n/	/n/
F_1	3.93	4.90	3.12
F_n	2.03	2.11	0.90

Tableau 3: rapports de variances (SPONT/LAB) des distributions de fréquences formantiques par consonne (un rapport supérieur à 1 indique une plus grande dispersion en parole spontanée).

Ce rapport est le résultat du quotient de la variance provenant des échantillons de parole spontanée par la variance provenant des échantillons de parole de laboratoire. Un coefficient unitaire indiquerait donc une stricte égalité de variance entre les deux styles de parole; un coefficient supérieur à l'unité manifestant une supériorité de la variance dans le cas des échantillons de parole spontanée. Ces données confirment l'examen des figures et indiquent, globalement, une supériorité nette de la dispersion dans le cas de la parole spontanée. Tous les rapports sont en effet supérieurs à 1, excepté celui qui caractérise le formant de nasalité de la consonne /n/. De manière générale, le rapport de variance est plus élevé dans le cas du premier formant (valeurs s'étendant de 3.12 à 4.9) que dans le cas du formant de nasalité (valeurs de .9 à 2.11).

6. CONCLUSIONS

L'ensemble des travaux présentés ici suggèrent d'importantes variations des sons envisagés sous l'effet du style de parole. Ces recherches sont cependant trop marquées encore par leur caractère exploratoire pour que de larges généralisations puissent en être inférées. Elles montrent cependant le profit à tirer d'une approche multiforme de la problématique envisagée et suggèrent des pistes de travail futur. L'accroissement de la variabilité des valeurs fomantiques en parole spontanée, qui paraît ici omniprésent semble ainsi un phénomène à approfondir. L'inconstance -d'une langue à l'autre- d'un phénomène classiquement rapporté (la centralisation) mérite également un examen plus approfondi. Les recherches actuellement en cours dans nos institutions s'y attachent.

7. REFERENCES

HARMEGNIES,B., POCH, D. "Vowel reduction in spontaneous speech in Spanish", *Proceedings of the ESCA Workshop on Phonetics and Phonology of Speaking Styles: reduction and elaboration in speech communications*, Barcelona, 1991, 311-315.

LOBACZ, P., "Speech rate and vowel formants", *Speech Analyzis and Synthesis*, vol.4, 1976, 186-218.

NORD, L., "Vowel reduction-centralization or contextual assimilation", *Speech Communication Seminar*, Stockholm, preprint version, 1974.

NORD, L., "Acoustic studies of vowel reduction in Swedish", *STL-QPSR*, vol. 4, 1986, 19-36.

den OS, E., "Vowel reduction in Italian and Dutch", *PRIPU*, vol. 10, 2, 1985, 3-12.

POCH, D., HARMEGNIES, B., "Variations structurelles des systèmes vocaliques en français et espagnol sous l'effet du style de parole", *Actes du Deuxième Congrès français d'Acoustique*, sous presse.

19es J.E.P. - Bruxelles, 19 au 22 mai 1992.

ÉTIQUETAGE PROSODIQUE ASCENDANT D'UN ÉNONCÉ

LANGLAIS P., MÉLONI H., VAISSIÈRE J.

LABORATOIRE D'INFORMATIQUE UNIVERSITE D'AVIGNON

Résumé

Cet article présente une étude permettant de mettre en évidence – de manière automatique – des phénomènes prosodiques utilisables dans un système de reconnaissance de la parole. Notre champs d'analyse de la substance prosodique est limité à la fonction supra-segmentale des paramètres acoustiques de fréquence fondamentale et de durée vocalique. Le corpus utilisé est constitué d'une cinquantaine de phrases de la base de donnée BDSON lues avec un débit moyen par 2 locuteurs. Il s'agit d'établir et de quantifier les interactions entre les paramètres prosodiques et les unités linguistiques d'un énoncé (mots, groupes de mots, syntagmes, etc.). Les résultats proposés sont obtenus à partir d'une segmentation manuelle (étiquetage de BDSON) ou automatique des énoncés. Ces connaissances seront utilisées aussi bien pour l'identification ascendante de limites des unités d'une phrase que pour la vérification descendante d'hypothèses sur l'organisation structurelle de l'énoncé.

1- INTRODUCTION

Le problème de l'identification de phénomènes prosodiques pertinents et de leur mise en correspondance avec la structuration d'un énoncé est très complexe et n'a pas reçu de solution satisfaisante malgré les nombreux travaux qui lui ont été consacré (Di Cristo, 1975). Après l'enthousiasme initial pour ce type de recherches (Lea 1975 ; Rossi 1980 ; Méloni 1982) plusieurs auteurs dressent plus récemment un bilan pessimiste concernant l'eficacité du traitement de ces connaissances dans les systèmes de reconnaissance de la parole (Waibel, 1988 ; Vaissière, 1988).

Malgré les ambiguités des informations prosodiques et les difficultés rencontrées pour en mesurer les paramètres significatifs, ces connaissances – qui ont une fonction déterminante pour la segmentation de l'énoncé en unités linguistiques diverses – doivent être prises en compte dans un système de reconnaissance de la parole continue.

Afin de quantifier les corrélations entre certaines unités linguistiques et les paramètres prosodiques, nous avons effectué une étude concernant la fonction démarcative de certaines configurations d'indices (rapport des durées des noyaux vocaliques, variations de F0, etc.). Nous comparons les résultats obtenus avec notre technique de calcul automatique de la durée des noyaux vocaliques, et ceux produits à partir de l'étiquetage manuel de BDSON.

2- LES PARAMETRES ET LES INDICES ACOUSTIQUES

Les indices acoustiques caractérisant certains phénomènes prosodiques sont représentés au moyen d'étiquettes et évalués à partir des seuls paramètres de durée et de fréquence fondamentale.

2.1 - Traitement de la durée des noyaux vocaliques

Un processus ascendant de décodage acoustico-phonétique fournit la localisation des noyaux vocaliques avec un taux de réussite de plus de 99% (Méloni, 1991). L'évaluation de la durée de la voyelle est particulièrement délicate et dépend à la fois du phonème et de son environnement – qui ne sont pas connus en situation de reconnaissance. Pour résoudre partiellement cette difficulté sans ajustement contextuel, nous mesurons la tenue de chaque voyelle en distinguant dans le noyau vocalique la zone temporelle de plus grande stabilité. Ce calcul simple rend compte de manière assez précise des allongements significatifs des voyelles. La durée syllabique est plus difficile à mesurer de manière automatique dans certains contexte.

Une voyelle dont la durée de la partie stable exède de quelques centisecondes la moyenne des intervalles vocaliques sur l'énoncé est affectée d'une étiquette (notée "AL") caractérisant cette situation. Chaque voyelle dont la durée est plus grande que celle de ses voisines immédiates est désignée au moyen du symbole "ED" ; à l'unité la plus longue est attribuée l'étiquette "MD".

2.2 - Traitement de la fréquence fondamentale

Sur les zones vocaliques, nous calculons la fréquence fondamentale au moyen d'une technique d'AMDF qui donne des résultats très satisfaisants dans ce contexte. Pour chaque voyelle nous choisissons comme valeur représentative de la fréquence fondamentale celle mesurée sur la fin de la zone stable. Dans le cadre de la reconnaissance de la parole il est difficile d'effectuer des corrections contextuelles (intrinsèques et co-intrinsèques) des valeurs de la fréquence fondamentale.

Les étiquettes caractérisant la situation d'une voyelle sur la courbe de la fréquence fondamentale désignent 3 configurations : le maximum de F0 sur l'énoncé "MF0", le minimum de F0 sur l'énoncé "mF0" et l'émergence bilatérale de l'unité "EF0". De plus, la dynamique de la fréquence fondamentale sur l'énoncé est découpée en 4 niveaux, et les voyelles dont la F0 émerge d'au moins 1 niveau par rapport à ses voisines sont caractérisées par le symbole "EN". Le choix du nombre des niveaux et de l'intervalle de variation de la fréquence fondamentale (limité à la phrase) résulte de nos expérimentations et de travaux déjà effectués sur ce problème (Di Cristo, 1981; Caelen, 1991).

3- ANALYSE DES CORRELATIONS ENTRE INDICES ET UNITES LINGUISTIQUES

Notre étude a pour but de mettre en relief les fonctions linguistiques d'indices prosodiques afin de déterminer un ensemble de règles utilisables dans un système de reconnaissance de la parole. Nous disposons pour cela d'un étiquetage (manuel et automatique) de chaque phrase prononcée ainsi que de sa décomposition en unités syntaxiques.

Une analyse quantitative des composantes du corpus (nombre d'unités linguistiques d'un type donné) permet de mesurer l'intervalle de confiance des résultats. Pour chaque type d'unité linguistique, on évalue le nombre d'étiquettes ou d'associations d'étiquettes qui affectent la première et la dernière voyelle de l'unité.

L'étude de 44 phrases phonétiquement équilibrées du corpus BDSON montre que la distribution des étiquettes n'est pas aléatoire. La figure 3 présente à cet effet, une table des correspondances entre des configurations d'étiquettes intéressantes et des unités structurelles simples. Il apparaît que les marques prosodiques se manifestent principalement en fin de mot et plus précisément sur les mots lexicaux. Le choix d'un corpus comportant des structures syntaxiques diverses implique des taux de corrélation moyens ; l'étude d'un sous-ensemble de ces phrases entraîne des taux nettement plus élevés que nous n'avons pas jugé utile de préciser.

Par ailleurs, ce travail a également fait apparaître que l'analyse du même corpus à partir d'une segmentation automatique n'entraîne pas de perte d'information significative. Les figures 1 et 2 montrent les divers taux de correspondance entre un groupe d'étiquettes et la fin des mots et des groupes fonctionnels des phrases. Il n'apparaît pas de détérioration notable des résultats.

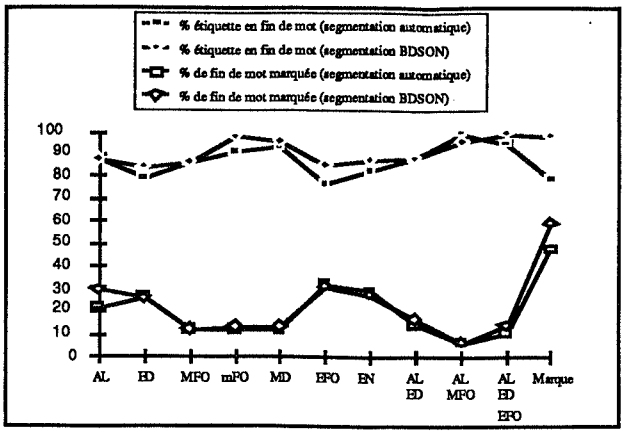


Figure 1 : pourcentages de corrélation entre des configurations d'étiquettes prosodiques données et les fins de mots.

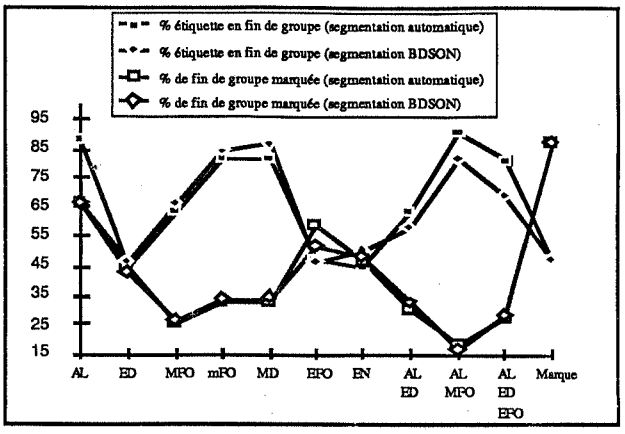


Figure 2 : pourcentages de corrélation entre des configurations d'étiquettes prosodiques données et les fins de groupes fonctionnels.

4- CONCLUSION

Au terme de cette étude, nous disposons d'un système d'étiquetage prosodique entièrement automatique capable de mettre en évidence dans une phrase, les sites de certains phénomènes prosodiques. Ce système ouvert permet l'intégration d'autres indices comme l'émergence de la fréquence fondamentale sur les syllabes au dessus d'une ligne de déclinaison.

Les résultats nous autorisent à penser qu'il est possible d'intégrer les connaissances prosodiques ainsi acquises dans un système de reconnaissance de la parole continue, tant dans une phase ascendante (émission d'hypothèses sur les fins de constituants) que dans une phase descendante de vérification. Une prochaine étape de nos travaux consistera donc en la réalisation d'une telle liaison dans le cadre d'une application limitée à des phrases de structures grammaticales régulières.

		AL	ED	MFO	mFO	MD	EFO	EN	AL + ED	AL + MFO	AL ED EFO	Marque
		111	102	44	44	44	123	105	62	22	45	204
Fin de Phrase	44	35			38	16						40
Fin de Sujet	44	20	23	18	1	16	29	27	17	11	16	32
Fin de Groupe Verbal	44	29	15	3	19	11	18	17	10	2	9	40
Fin de Complément	33	24	9	8	17	11	10	9	9	5	6	25
Début de Phrase	44	5		1	2	1						5
Début de Sujet	44	2			1	1						3
Début de Gpe Verbal	44	7	7	1	1	3	5	2	5	1	2	12
Début de Complément	33	5	3	2	1	1			1			7
Fin de Mots	339	97	86	38	43	42	104	92	55	21	45	201
Fin de Mots Gramm.	172	9	13	4	3	5	14	8	5	2	3	31
Fin de Mots Lexicaux	167	88	73	34	40	37	90	84	50	19	42	170
Fin de Groupe	111	73	47	29	37	38	57	53	36	18	31	97
Début de Groupe	111	14	10	3	3	5	5	2	6	1	2	22

Figure 3 : Corrélations entre certaines positions intéressantes et les configurations d'étiquettes prosodiques étudiées , à partir de l'étiquetage BDSON. La première ligne de valeurs représente le nombre total d'étiquettes positionnées sur les phrases du corpus. La première colonne précise le nombre d'occurrences de chaque position. La colonne **marque** indique un allongement de durée ou une émergence quelconque.

BIBLIOGRAPHIE

G. Caelen (1991), "Stratégies des locuteurs et consignes de lecture d'un texte : analyse des interactions entre modèles syntaxiques, sémantiques, pragmatiques et paramètres prosodiques", *Thèse d'Etat*, Aix-en-Provence.

A. Di Cristo (1975), "Soixante-dix ans de recherches en prosodie", *Editions de l'Université de Provence.*

A. Di Cristo (1981), "Aspect phonétique et phonologique des éléments prosodiques", *Modèles linguistiques*, Tome III, Fascicule 2, pp. 24-83.

W.A. Lea, M.F. Medress, T.E. Skinner (1975), "A prosodically guided speech understanding strategy", *Proc. IEEE ICASSP*, Vol. 1, pp. 30-38.

H. Méloni, J. Guizol (1982), "Utilisation de paramètres prosodiques dans un système de reconnaissance automatique de la parole continue", *Actes du Séminaire Prosodie et Reconnaissance Automatique de la Parole*, Aix-en-Provence, pp. 93-120.

H. Méloni (1983), "Traitement des contraintes linguistiques en reconnaissance de la parole", *Revue Techniques et Sciences Informatiques*, Vol. 2, n° 5, pp. 349-363.

H. Méloni, P. Gilles (1991), "Décodage acoustico-phonétique ascendant", *Revue Traitement du Signal*, Vol. 8, n° 2, pp. 107-114.

M. Rossi, A. Di Cristo (1980), "Un modèle de détection automatique des frontières intonatives et syntaxiques", *11èmes JEP*, Strasbourg, pp. 217-238.

J. Vaissière (1984), "PROSEIDON : Automatic detection of prosody cues in continuous speech", 13èmes JEP, Bruxelles, pp.189-190.

J. Vaissière (1988), "The use of prosodic parameters in automatic speech recognition", in H. Niemann, M. Lang & G. Sagerer (Eds.), Recent Advances in Speech Understanding and Dialog Systems, NATO ASI Series, Berlin : Springer-Verlag, pp. 71-99.

A. Waibel (1988), "Prosody and speech recognition", Londres : Pitman.

19es J.E.P. - Bruxelles, 19 au 22 mai 1992.

Une Architecture Connexionniste Modulaire pour l'Identification Automatique du Locuteur

Younès BENNANI & Patrick GALLINARI

Laboratoire de Recherche en Informatique
Université de Paris-Sud, Centre d'Orsay Bât. 490
U.A. 410 C.N.R.S. 91405 Orsay FRANCE
e-mail : younes@lri.lri.fr

Résumé

Nous présentons un système connexionniste pour l'identification du locuteur en mode indépendant du texte. Nous avons mis au point une architecture composée de plusieurs modules connexionnistes qui coopèrent pour l'identification. Le système est composé d'un détecteur de typologie et d'un ensemble de modules experts. Chaque module expert du système est conçu pour la discrimination entre les locuteurs de même typologie. Le module de détection des typologies intervient dans la décision finale par multiplication de son score par ceux des modules experts. Le système a été testé sur une population de 102 locuteurs extraite de la base DARPA-TIMIT. Une parfaite identification a été observée, soit un interval de confiance à 95% de [99.9%,100%] avec une précision de 0.1%. Les performances de notre système ont été comparées avec celles d'un système basé sur l'approche des modèles auto-régressifs vectoriels.

1 INTRODUCTION

Nous présentons dans ce papier un système pour l'identification du locuteur qui est opérationnel sur un grand nombre de locuteurs et que nous avons testé sur une partie de la base TIMIT. Il s'agit d'un système complexe qui réalise une intégration des modules d'extraction de caractéristiques et de classification.
Pour le mettre au point, nous avons utilisé une démarche qui est nouvelle dans le domaine connexionniste. Il s'agit de la décomposition d'une tâche principale en un ensemble de sous tâches dont la résolution permet celle du problème initial. Cela nous a amené à concevoir des systèmes modulaires où différents réseaux coopèrent à une même tâche. Dans ces systèmes, la puissance de calcul n'est pas apportée par un réseau unique possédant un grand nombre d'unités non linéaires, mais par la coopération des différents modules qui coopèrent à la tâche.

Cela permet de gérer des réseaux de petite taille, et, quand la complexité du problème s'accroît, de rajouter des modules qui restent assez simples. Dans l'approche non modulaire, la seule réponse à l'accroissement de la difficulté d'une tâche est l'utilisation de modèles de plus en plus complexes, dont le nombre de paramètres devient tel qu'il est très difficile de réaliser l'apprentissage du système.

Cette démarche est totalement générale et peut être utilisée pour résoudre des problèmes issus de domaines complètement différents. En agissant de la sorte, on ne fait que transposer au niveau des modules la philosophie connexionniste et l'on crée en quelque sorte des réseaux de réseaux. Nous verrons plus en détail par la suite quels sont les autres avantages de cette approche.

Les principales caractéristiques de notre système sont les suivantes :

• un module connexionniste basé sur des architectures de type TDNN est utilisé pour l'extraction des caractéristiques (modélisation).

Ce système travaille directement sur le signal paramétré et réalise donc une extraction dynamique des caractéristiques.

• La prise en compte d'un grand nombre de locuteurs est rendue possible grâce à l' architecture modulaire. Celle-ci reflète la nature de la tâche à résoudre et prend en compte une typologie des locuteurs.

• le système fonctionne en mode indépendant du texte. L'apprentissage et le test ont été réalisés sur une base internationale qui est la base DARPA-TIMIT.

Les avantages de notre approche sont les suivants :

• L'extraction des caractéristiques et la classification se font par un même modèle et donc en même temps. Les deux étapes sont donc optimales l'une par rapport à l'autre, ce qui est rarement réalisé dans les systèmes classiques de reconnaissance.

• L'architecture utilisée pour l'extraction des caractéristiques tient compte de la nature dynamique du signal de la parole. Cette architecture est capable de représenter les relations temporelles entre les vecteurs acoustiques.

• Les modules utilisés sont simples et de taille réduite grâce à l'architecture modulaire.

• Le système en phase d'identification fonctionne en "temps réel" ce qui n'est pas le cas des autres approches en IAL.

• nous avon mis en oeuvre le système sur une population de taille statistiquement représentative (102 locuteurs : 33 femmes et 69 hommes). Cela en fait un des systèmes les plus importants qui ait été testé. La philosophie suivant laquelle il est construit permet de l'étendre sans peine.

Nous présenterons tout d'abord une brève description de la base de données utilisée pour l'évaluation du système. Nous décrivons ensuite les techniques de prétraitements et d'analyse effectués sur le signal de parole. Une nouvelle architecture de réseau TDNN (STDNN) sera présentée. Après une description de l'étude expérimentale, nous présenterons les résultats puis une comparaison des performances de notre système avec une autre technique (MAV).

2 CONSIDÉRATIONS GÉNÉRALES

2.1 Description de la Base de Données

C'est une base américaine composée de 420 locuteurs appartenant à 8 dialectes [Fisher, 87]. Chaque locuteur prononce 10 phrases : 5 phonétiquement riches il s'agit des phrases (SX), 3 phrases représentant des phrases naturelles (SI) et 2 phrases (SA) reflétant le dialecte des locuteurs. Ces deux dernières phrases sont les mêmes pour tous les locuteurs.

Nous avons choisi les 5 phrases phonétiquement équilibrées pour l'apprentissage de notre système et les 5 autres phrases pour tester le système. Celles-ci sont différentes pour tous les locuteurs. Il s'agit donc bien d'un système indépendant du texte.

2.2 Prétraitement et Analyse du Signal Acoustique

Les signaux ont été enregistrés sur un disque CDROM de très bonne qualité. L'échantillonnage est de 16 kHz codé sur 16 bits.

Une analyse LPC à l'ordre 16 a été réalisée. Chaque bloc d'analyse est pondéré par une fenêtre de Hamming de longueur 25,6 msec. La corrélation du signal est calculée toutes les 10 msec, pour 16 coefficients. Une pré-emphase de 94% est effectuée sur le signal numérique.

On conserve ensuite les coefficients cepstraux (LPCC) calculés à partir des coefficients de prédiction.

Après la paramétrisation LPCC, chaque phrase est représentée par une matrice 16xN, où N représente le nombre de trames de la phrase.

Cette paramétrisation a été largement utilisée en RAL. Une des raisons est que les paramètres basés sur l'analyse par prédiction linéaire contiennent des informations relatives aux formants, à l'onde glottale et à la radiance des lèvres [Atal, 74].

Nous avons utilisé une partie importante de la base TIMIT correspondant à une centaine de locuteurs.

3 UNE ARCHITECTURE CONNEXIONNISTE MODULAIRE POUR L'IAL

3.1 Décomposition de la tâche d'IAL par Modularité

L'IAL est en général bien plus complexe en mode indépendant du texte qu'en mode dépendant. Toutefois contrairement à la vérification, les applications de l'IAL nécessitent la plupart du temps une reconnaissance indépendante du texte. On ne peut plus alors utiliser des techniques simples d'appariement de références. L'apprentissage demande un très grand nombre de données, par exemple de l'ordre de quelques dizaines de secondes, et le test demande plusieurs secondes de signal. Le temps d'apprentissage pour un tel système croît bien sûr avec le nombre de locuteurs à identifier. Dans les systèmes non discriminants, il croît en général de façon linéaire avec ce nombre, dans les systèmes discriminants, la croissance est plutôt de type exponentiel. Remarquons qu'il n'en est pas forcément de même lors de la phase de reconnaissance.

En ce qui concerne l'extraction de caractéristiques à partir du signal paramétré, il est possible d'utiliser une représentation mettant en évidence les caractéristiques à long terme du signal comme nous l'avons fait dans notre premier système. Il faut toutefois remarquer que, à cause de l'augmentation de la durée du signal d'apprentissage, qui seule permet d'assurer l'indépendance du vocabulaire, cette étape est bien plus longue à réaliser. De plus, lors de la phase de reconnaissance, il faut analyser toute l'émission avant de la soumettre au système. Nous avons donc décidé de travailler en dynamique sur le signal, ce qui permet de plus de conserver la dimension temporelle.

Nous avons voulu conserver une propriété très importante des réseaux connexionnistes qui est la possibilité de prendre en compte la structure inter-classe des données et donc de réaliser un apprentissage discriminant. Toutefois, comme nous l'avons mentionné, cette prise en compte augmente de façon plus que linéaire la durée de l'apprentissage en fonction du

nombre de locuteurs. Il devient donc rapidement impossible de mettre en oeuvre des réseaux pour l'identification quand le nombre de locuteurs augmente. Nous avons eu l'occasion de comparer le temps d'apprentissage pour l'identification avec ceux nécessaires pour résoudre des problèmes de taille similaire pris dans d'autres domaines. Les temps pour l'identification sont considérablement plus longs, ce qui illustre la difficulté du problème.

Une réponse à cette difficulté est de décomposer le problème en problèmes plus simples à résoudre et d'affecter à chaque sous-problème un réseau. Ces réseaux sont assemblés pour constituer une architecture modulaire et coopèrent au problème global. Le nombre de modules croît avec la complexité du problème à résoudre, mais chacun d'entre eux reste de taille limitée et est donc facile à entraîner. Par comparaison avec une architecture unique, le temps d'apprentissage est extrêmement réduit et croît de façon quasi linéaire avec le nombre de locuteurs. Bien sûr, la décomposition de la tâche doit refléter la structure du problème. Si cela est le cas, la structure modulaire correspondante incorporera cette connaissance a priori que nous avons sur la structure du problème. De façon optimale, cette décomposition d'une tâche en sous tâches pourrait être faite de façon automatique et faire ainsi partie de l'apprentissage du système. En l'état actuel de nos connaissances, seules de timides propositions ont été réalisées dans ce sens et nous n'avons pas voulu nous y lancer pour une véritable application. Toutefois, pour un problème donné, la décomposition peut être réalisée en utilisant des heuristiques qui mettent en oeuvre des règles de bon sens.

Dans la première version du système [Bennani & Gallinari, 91], nous avons testé la mise en oeuvre d'une architecture modulaire simple. L'idée même de faire coopérer différents modules connexionnistes était d'ailleurs toute neuve à l'époque et rien n'assurait sa faisabilité.

Durant notre travail sur l'IAL, nous avons remarqué que certaines caractéristiques permettent de former des classes homogènes et séparées. A titre d'exemple, la période fondamentale moyenne (Fo moy) fournit deux classes bien distinctes, celle des femmes et celle des hommes. Cette connaissance a priori sur la tâche globale sera insérée dans le système pour réaliser deux modules séparés [Bennani & Gallinari, 91].

3.2 Incorporation de connaissances dans l'architecture du système

A l'intérieur de la classe "femmes" ou de la classe "hommes" nous avons distingué également plusieurs sous-classes. Celles-ci regroupent les locuteurs ou locutrices dont les caractéristiques vocales se ressemblent le plus. Afin d'automatiser la détermination des typologies des locuteurs par un modèle connexionniste nous avons étiqueté la base d'apprentissage par les typologies des locuteurs au lieu de l'identité. L'étiquetage a été effectué au moyen d'une technique de "clustering" de type k-moyennes suivie d'un vote majoritaire. En effet, les vecteurs acoustiques d'apprentissage (phrases SX) modélisant les voix des locuteurs ont été regroupés en un ensemble de classes homogènes à l'aide de l'algorithme non supervisé K-means. Ces classes représentent donc les différentes typologies de la population étudiée. Après ce regroupement des vecteurs acoustiques, nous procédons à un vote majoritaire par locuteur et par classe. On affecte un locuteur à la typologie (classe) dans laquelle il y a la majorité de ses vecteurs acoustiques.

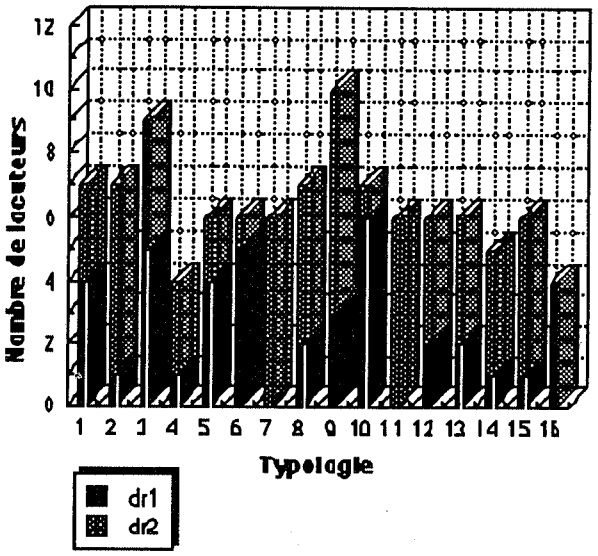


Figure 1 : Distribution des locuteurs par typologies et par dialectes

Après ce regroupement des locuteurs en classes homogènes, nous n'avons rencontré aucune exception de femmes ayant des typologies d'hommes ou d'hommes ayant des typologies de femmes. Par contre dans la plupart des groupes un mélange entre les deux dialectes a été observé avec un dialecte dominant.

Nous avons alors mis en oeuvre un système multi-modulaire pour l'IAL utilisant ces connaissances a priori sur la tâche.

3.3 Architecture du Système Connexionniste

Le système est composé d'un détecteur de typologie et d'un ensemble de modules experts. Chaque module expert du système est conçu pour la discrimination entre les locuteurs de même typologie (les confusions entre les locuteurs de même typologie sont plus importantes que celles entre les locuteurs de typologies différentes). Le module spécialisé en détection des typologies des locuteurs joue le rôle d'aiguilleur de l'information. L'architecture de ce nouveau système peut être vue sous deux formes [Bennani, 92]. Un premier cas où le module de détection des typologies intervient dans la décision finale par multiplication de son score par ceux des modules experts. Le deuxième cas où le module de détection des typologies sert à orienter le message vers le

module expert approprié. Il faut cependant noter que le temps d'identification dans le deuxième cas est nettement inférieur à celui du premier cas. En effet la pré-sélection du module expert concerné évite un calcul long et fastidieux par les autres modules experts et réduit considérablement le temps de calcul. Mais une erreur commise pendant la détection de la typologie sera pénalisée par les modules experts avec la première architecture, ce qui n'est pas le cas pour la deuxième. Pour nos simulations, nous n'avons pas remarqué de différence de performances entre les deux types d'architectures, sachant que le module détecteur de typologie permettait une parfaite distinction entre les typologies des locuteurs.

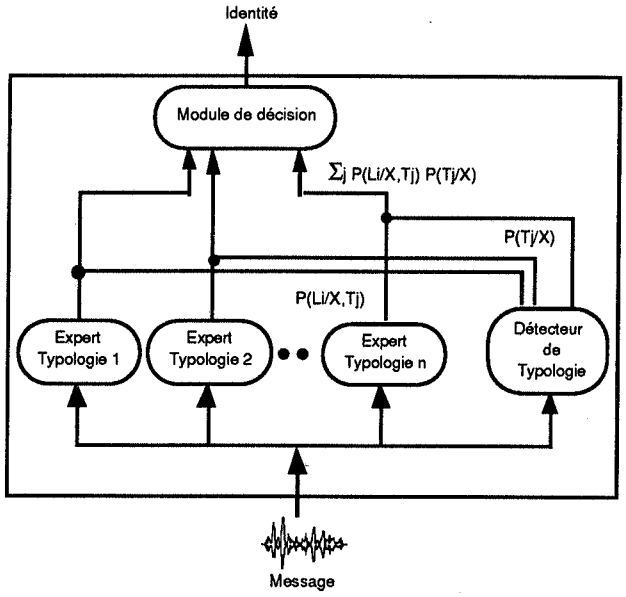


Figure 2 : Architecture du Système Connexionniste Multi-Modulaire

3.3.1 Composants du Système

En reconnaissance automatique de la parole, toutes les données en entrée sont indicées par le temps. Ne pas en tenir compte revient à imposer au système de reconnaissance une charge alourdie.

L'idée de prendre en compte la notion du temps dans un réseau multi-couches a été initialement introduite dans [Lang & Hinton, 88], sous le nom de TDNN (Time Delay Neural Network).

Les composants de l'architecture de notre système d'IAL, sont constitués de modèles connexionnistes du type TDNN. La taille de l'entrée de ces réseaux est fixe comme dans la plupart des modèles connexionnistes. Ceci pose un problème avec les données de parole où les phrases n'ont pas la même taille.
Dans notre ancien système [Bennani et al., 90] la modélisation a été effectuée par une technique statistique. Cette technique nous a permis de représenter chaque phrase par le vecteur spectre moyen et le premier vecteur propre de la matrice de covariance. Cette technique de modélisation est indépendante de la taille de la phrase et produit des entrées de taille fixe pour le classifieur connexionniste.

Ici, nous procédons différemment, en faisant glisser une fenêtre de taille fixe sur chaque phrase. Chaque position de la fenêtre fournira une entrée au système. L'ensemble des données concernant une phrase sera caractéristique d'un locuteur. Plus précisément nous procédons de la façon suivante :

Nous divisons chaque phrase (les paramètres LPCC issus du signal) en une suite de fenêtres. Chaque fenêtre est composée de 25 vecteurs spectraux ou trames avec un recouvrement de 20 trames. Ces fenêtres constituent les entrées des trois modules décrits précédemment.

3.3.2 Phase d'Apprentissage du système

Pendant la phase d'apprentissage, les fenêtres successives obtenues par cette décomposition de la phrase serviront à la détermination des paramètres du système. L'ensemble d'apprentissage d'un module sera constitué par l'ensemble des fenêtres ainsi découpées sur les phrases d'apprentissage. Ainsi, quand on présente une fenêtre de 25 trames au module M_i, on modifie les paramètres de M_i de façon à se rapprocher de la réponse désirée.

3.3.3 Phase d'Identification

Pour l'identification, toutes les trames composant la phrase sont successivement présentées au système sous forme de fenêtres de 25 vecteurs acoustiques. A chaque présentation d'une fenêtre le système produit une réponse (le nom du locuteur le plus probable). Les activations successives du système sont ensuite additionnées sur la longueur de la phrase. La décision finale est faite sur le locuteur possédant le plus grand score.

Nous appellerons ce type de TDNN : STDNN (pour Shift-TDNN). La figure 3 explique le principe du calcul des activations dans le temps et par la suite les activations finales dans un STDNN.

Nous avons remarqué dans nos expériences qu'après apprentissage, il suffit d'une suite de trois fenêtres pour une parfaite identification. Si l'on compare avec les systèmes existants, ces trois fenêtres correspondent à une émission extrêmement courte (moins d'une seconde) pour l'identification en mode indépendant du texte. La plupart des systèmes utilisent des émissions de l'ordre de plusieurs secondes [O'Shaugnessy, 87], pour la base TIMIT, [Rudosi & Zahorian, 91] par exemple ont utilisé 8 secondes d'émission pour la reconnaissance.

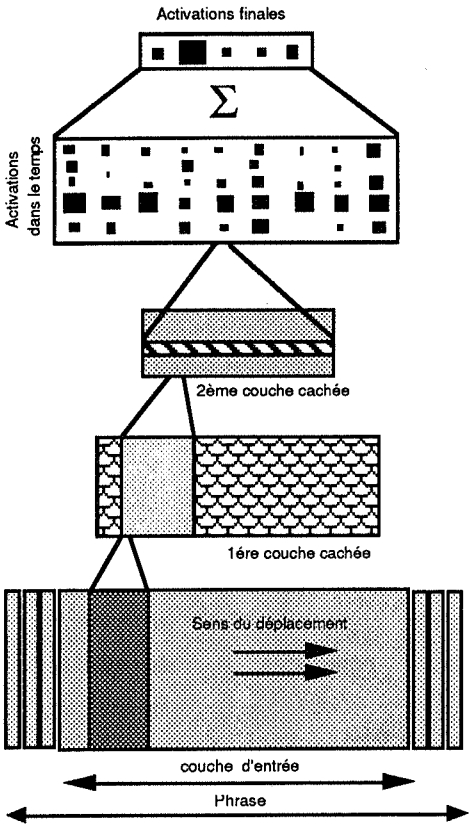


Figure 3 : Architecture STDNN

4 MODELES AUTO-RÉGRESSIFS VECTORIELS (MAV) POUR L'IAL

Cette technique a été utilisée pour la première fois pour l'identification du locuteur par Y.GRENIER [Grenier, 80].

Celui-ci considérait que *"le modèle auto-régressif vectoriel calculé sur la voix d'un locuteur modélise les capacités articulatoires du locuteur, du moins en première approximation "*.

A partir de la concaténation des 5 phrases d'apprentissage, un modèle est construit. Chaque locuteur possède donc un modèle. En phase d'identification, on calcule un modèle du locuteur à partir de la phrase destinée au test.

La décision finale consiste à comparer le modèle de la phrase test avec la totalité des modèles références des locuteurs et choisir comme identité l'identité du locuteur dont le modèle référence est le plus proche du test. L'ordre du modèle est difficile à déterminer, nous l'avons fixé à 2, vu le nombre disponible des données.

La description de cette technique est donnée dans [Bennani, 92].

5 RÉSULTATS ET COMPARAISON

Les résultats de la table 1 montrent la supériorité de l'approche connexionniste multi-modulaire par rapport à la technique MAV.

Approche	*Score*	*Précision*	*Int. Conf. à 95 %*
STDNN	100 %	± 0.1 %	[99.9 %, 100 %]
MAV	97.5 %	± 6.1%	[91.4 % , 95.6 %]

***Table 1** : Comparaison des deux approches : Connexionniste et MAV*

On peut remarquer une très grande différence entre les tailles des intervalles de confiance pour les deux techniques. Ceci est dû au fait que les modèles MAV demandent la totalité de la phrase test pour estimer le modèle êt effectuer l'identification, ce qui rend la taille de l'ensemble de test égal au nombre de phrases tests par locuteur (5 tests par locuteur). Par contre l'approche connexionniste se contente de deux ou trois fenêtres de 25 trames pour une parfaite identification, ceci offre un grand nombre de test (≈100 tests par locuteur). L'intervalle de confiance est calculé en fonction du nombre de test et du taux de reconnaissance (la méthode de calcul de cet intervalle est donnée en annexe). Par conséquent, un grand nombre de tests donne un "petit" intervalle de confiance et un petit nombre de tests donne un "grand" intervalle de confiance. La taille de l'intervalle de confiance reflète en quelque sorte la robustesse du système et la validité de ses performances.

L'avantage de notre approche connexionniste comparée à d'autres techniques réside essentiellement dans :

- le temps d'identification, moins d'une seconde, on peut donc dire qu'on a une identification en temps réel,
- la courte durée du signal de parole nécessaire pour l'identification.
- la validité de performances du système.

6 INTERPRÉTATION DES FONCTIONS CALCULÉES PAR LES RÉSEAUX MODULAIRES

Nous donnons rapidement ci dessous une interprétation probabiliste du fonctionnement de nos réseaux modulaires. Désignons par L_i (i = 1 .. m) l'identité du locuteur i, et T_j (j =1 .. n) la jème typologie de la population étudiée.

Si l'on considère par exemple l'architecture de la figure 2, le réseau détecteur de typologie va calculer, pour une forme acoustique X, une approximation de $P(T_j /X)$ pour $j = 1..n$. Les réseaux experts qui sont entraînés indépendamment les uns des autres vont calculer une approximation de $P(L_i /X,T_j)$ pour le réseau j (On considère que dans la fonction calculée, $P(L_i /X,T_j) = 0$ si L_i n'est pas dans T_j).

L'opération réalisée en sortie du système est pour la sortie i une approximation de :

$$\sum_{j=1}^{n} P(L_i /X,T_j)\ P(T_j /X) \qquad (1)$$

où l'on reconnait l'expression de $P(L_i /X)$.

Le système de la figure 2 fournit donc à travers une architecture modulaire une estimation de $P(L_i /X)$ [Bennani, 92] au moyen des approximation successives des probabilités figurant dans l'expression (1). Il présente l'inconvénient suivant : pour les valeurs calculées en sortie, les erreurs d'approximation des réseaux successifs sont combinées.

On peut du moins en théorie entraîner le même réseau à prédire directement en sortie $P(L_i /X)$. On obtiendra dans ce cas une estimation de $P(L_i /X)$ qui sera la meilleure approximation au sens des moindres carrés pour l'architecture utilisée et sera donc différente de celle fournie par (1).

7 ALGORITHME D'APPRENTISSAGE GLOBAL

Pour réaliser un apprentissage global du système de la figure 2, on peut utiliser un algorithme de gradient adaptatif similaire à la rétro-propagation. La seule différence entre l'architecture classique d'un MLP et celle du système de la figure 2 est la présence dans ce dernier de connexions multiplicatives entre le module de typologie et les modules experts.

Un algorithme permettant la mise à jour de connexions multiplicatives a par exemple été proposé par [Hampshire & Waibel, 90]. La dérivation en est très simple et est similaire à celle de la rétro-propagation.

Une bonne solution consiste sans doute à initialiser le système avec les algorithmes locaux et ensuite à tenter de l'améliorer avec l'algorithme global. Cette approche s'est déjà révélée fructueuse pour d'autres problèmes et d'autres systèmes.

8 DISCUSSION

Nous avons mené un ensemble d'expériences qui a montré la validité de l'approche connexionniste pour IAL en mode indépendant du texte. Ces modèles permettent d'extraire les traits nécessaires pour la discrimination entre les locuteurs et en même temps d'effectuer la classification. Notre approche modulaire est très générale, elle nous a permis d'insérer des connaissances a priori sur le problème et de résoudre la tâche globale en la décomposant en sous-tâches de difficulté moins importante. Notre approche modulaire réduit considérablement le temps de calcul dans le système et autorise une identification en temps réel.

Enfin la comparaison de notre système avec d'autres techniques différentes a mis en évidence la puissance et l'intérêt de l'approche connexionniste en IAL. Cette comparaison nous a donné l'idée de faire coopérer les différentes techniques entre elles afin de combiner leurs mérites respectifs et concevoir des systèmes hybrides.

9 REFERENCES

Artieres T., Bennani Y., Gallinari P. (1991)
"Connectionist and Conventional Models for Free-Text Talker Identification Tasks", proc. of Neuro-Nimes 91, France.

Atal B.S. (1974)
"Effectiveness of LPC characteristics of the speech wave for A.S.I and A.S.V", JASA-Vol.55.1974.

Bennani Y. (1992)
"Approches Connexionnistes Pour la Reconnaissance Automatique du Locuteur : Modélisation & Identification", Ph.D. thesis, Université de Paris-Sud, janvier '92.

Bennani Y., Fogelman F., Gallinari P. (1990)
"Text-Dependent Speaker Identification Using Learning Vector Quantization", proc. of INNC, July Paris, FRANCE.

Bennani Y. & Gallinari P. (1991)
"On The Use Of TDNN- Extracted Features Information In Talker Identification", proc. of ICASSP, S6.5, Toronto, Canada.

Fisher W., Zue V.,Bernstein J., Pallett D. (1987)
"An Acoustic-Phonetic Data Base", J. Acoust. Soc. Amer. Suppl. (A), 81, S92.

Furui S. (1981)
"Cepstral Analysis technique for automatic speaker verification" IEEE Trans. on ASSP, Vol. 29, N° 2.

Grenier Y. (1980)
"Utilisation de la Prédiction Linéaire en Reconnaissance et Adaptation au Locuteur", proc. of XIème JEP, Strasbourg, pp. 163-171.

Hampshire J.B. & Waibel A.H. (1990)
"The Meta-Pi Network: Connectionist Rapid Adaptation For High-Performance Multi-Speaker Phoneme Recognition", proc. of ICASSP, S3.9, NM, USA.

Jacobs R.A.(1990)
"Task Decomposition Through Competition in a Modular Connectionist Architecture", Ph.D. thesis, University of Massachussets at Amherst.

Oglesby J. & Mason J.S. (1990)
"Optimisation of Neural Models for Speaker Identification", proc. of ICASSP, S5.1, NM, USA.

O'Shaugnessy D. (1986)
"Speaker Recognition", IEEE ASSP Magazine, Vol. 3, pp. 4-17.

Rosenberg A.E., Lee C.H., Soong F. (1991)
"Sub-Word Unit Talker Verification Using Hidden Markov Models", proc. of ICASSP, S5.3, NM, USA.

Rudasi L. & Zahorian S.A. (1991)
"Text-Independent Talker Identification With Neural Networks", proc. of ICASSP, S6.6, Toronto, Canada.

Zheng Y.C. & Yuan B.Z. (1988)
"Text-Dependant Speaker Identification Using Circular Hidden Markov Models". Proc. of ICASSP, S13.3, pp. 580-582.

Liste des auteurs

Christian ABRY, *ICP - Université Stendhal*
Emmanuelle ABSIL, *Université de l'Etat de Mons*
Nour-Edine ACHAB, *ICP - Université Stendhal*
Gilles ADDA, *LIMSI - CNRS - Orsay*
Lourdes AGUILAR, *Université de Barcelone*
O. AL DAKKAK, *ISSAT - Damas*
F. ANDRY, *CAP GEMINI INNOVATION*
Yolande ANGLADE, *Université de Nancy I*
Danièle ARCHAMBAULT, *INRS-Télécommunications*
Véronique AUBERGE, *ICP - Université Stendhal*
Denis AUTESSERRE, *Université de Provence I*
Pierre BADIN, *ICP - INP Grenoble*
Gérard BAILLY, *ICP-INP Grenoble*
P. BARBOSA, *ICP-Université Stendhal*
Claude BARRAS, *Université Paris VI*
Frédéric BEAUGENDRE, *LIMSI - CNRS - Orsay*
F. BECHET, *Université d'Avignon*
Rabia BELRHALI, *ICP - Université Stendhal*
Zina BEN MILED, *IRSIT - Tunis*
Younes BENNANI, *Université de Paris-Sud*
Christian BENOIT, *ICP - Université Stendhal*
Abdelkader BETARI, *Faculté des Sciences de Luminy*
E. BILANGE, *CAP GEMINI Innovation*
Louis-Jean BOE, *ICP - Université Stendhal*
Denys BOITEAU, *C.N.E.T - Lannion*
Jean-François BONASTRE, *Université d'Avignon*
Anne BONNEAU - CHAREAU, *Université de Nancy I*
Stéphane BORNERAND, *BULL*
Bachir BOUDRAA, *U.S.T.H.B. - Alger*
Malika BOUDRAA, *U.S.T.H.B. - Alger*
Marie-Luce BOURGUET, *ICP - INP Grenoble*
C. BOURJOT, *Université de Nancy I*
Hervé BOURLARD, *Lernout & Hauspie Speech Products*
A. BOYER, *Université de Nancy I*
Abdelfattah BRAHAM, *IRSIT - Tunis*
Rémy BULOT, *Faculté des Sciences de Luminy*
Jean CAELEN, *ICP - INP Grenoble*
Geneviève CAELEN, *ICP - INP Grenoble*
Jean-Claude CAEROU, *ICP - INP Grenoble*
J. CAMPS, *ICP - INP Grenoble*
Marie-José CARATY, *Université Paris VI*
René CARRE, *ENST - Télécom Paris*
Michel CARTIER, *C.N.E.T - Lannion*
Eric CASTELLI, *ICP - INP Grenoble*
Marie Agnès CATHIARD, *Université Pierre Mendès-France*
Christian CHANARD, *Université Nationale de Côte d'Ivoire*
François CHARPENTIER, *CAP GEMINI INNOVATION*
F. CHARPILLET, *Université de Nancy I*
K. CHOUKRI, *CAP GEMINY INNOVATION*
P. COMBESCURE, *C.N.E.T - Lannion*
Daniel COTTO, *Université Paul Sabatier*
Christophe D'ALESSANDRO , *LIMSI - CNRS - Orsay*
Delphine DAHAN, *CNRS - Paris*
Martine DE CALMES, *Université Paul Sabatier*
Raoul DE GUCHTENEERE, *Université Libre de Bruxelles*
Paul DELEGLISE, *ENST - Télécom Paris*
Bernard DELYON, *IRISA - CNRS - Rennes*
Didier DEMOLIN, *Université Libre de Bruxelles*
Maria Gabriella DI BENEDETTO, *Université La Sapienza*
H. DIA, *ICP - Université Stendhal*
Amar DJERADI, *USTHB - Alger*
Jean-Marc DOLMAZON, *ICP - INP Grenoble*
Marc DOMINICY, *Université Libre de Bruxelles*
Fabrice DUERMAEL, *Université de Nancy I*
P. DUPONT, *Philips Research Laboratory*
Abdelhamid EL BADMOUSSI, *ICP - INP Grenoble*
F. EMERARD, *C.N.E.T - Lannion*
Pierre ESCUDIER, *ICP - INP Grenoble*
Robert ESPESSER, *Université de Provence*
Azarshid FARHAT, *Université Paul Sabatier*
G. FENG, *ICP - Université Stendhal*
Isabelle FERRANE, *Université Paul Sabatier*
Dominique FOHR, *Université de Nancy I*
Dominique FRANCOIS, *Université de Nancy I*
Patrick GALLINARI, *Université de Paris-Sud*
Martine GARNIER, *LIMSI - CNRS - Orsay*
F. GAVIGNET, *CAP GEMINI INNOVATION*
Claire GERARD, *CNRS - Paris*
Salem GHAZALI, *IRSIT - Tunis*
P. GILLES, *Université d'Avignon*
Sylvain GITTON, *C.N.E.T - Lannion*
Y. GONG, *Université de Nancy I*
Sophie GRAU-GOVEL, *LIMSI - CNRS - Orsay*
Isabelle GUAITELLA, *Université de Provence I*

Richard GUBRYNOWICZ, *Institut des Recherches Fondamentales de Technologie*

Bernard GUERIN, *ICP - INP Grenoble*

Jomini HABAILI HUSSEIN, *IRSIT - Tunis*

Bernard HARMEGNIES, *Université de l'Etat de Mons*

Jean-Paul HATON, *Université de Nancy I*

Daniel HOLENDER, *Université Libre de Bruxelles*

Jean-Marie HOMBERT, *Université Lumière Lyon 2*

Mohamed JEMNI, *IRSIT - Tunis*

Guéorgui JETCHEV, *Université de Sofia*

Paul JOSPA, *Université Libre de Bruxelles*

S. KITAZAWA, *Université de Shizuoka*

J. KLEIN, *Université de Nancy I*

Régine KOLINSKY, *Université Libre de Bruxelles*

J.-P. KOSTER, *Université de Trêves*

Rafael LABOISSIERE, *ICP - INP Grenoble*

Anne LACHERET DUFOUR, *Université de Caen*

Med-Tahar LALLOUACHE, *ICP - Université Stendhal*

Albert LANDERCY, *Université de l'Etat de Mons*

Philippe LANGLAIS, *Université d'Avignon*

Yves LAPRIE, *Université de Nancy I*

Jean-Luc LE FLOCH, *Université de Paris VI*

Laure LIBERT, *ICP - Université Stendhal*

A. LICHENE, *ICP - INP Grenoble*

Jean-Sylvain LIENARD, *LIMSI - CNRS - Orsay*

Mohamed Nabil LOKBANI, *C.N.E.T - Lannion*

Jesus MACHUCA, *Université de Barcelone*

Jean-Yves MAGADUR, *CAP GEMINI INNOVATION*

A. MARCHAL, *Université de Provence*

Pierre MARQUIS, *Université de Nancy I*

J. MARTI, *Université Ramon Llull*

Nicole MARTIN, *Université de l'Etat de Mons*

Gemma MARTINEZ, *Université de Barcelone*

Laurent MAUUARY, *C.N.E.T - Lannion*

Odile MELLA, *Université de Nancy 1*

Henri MELONI, *Université d'Avignon*

C. MEUNIER, *Université de Provence*

Tayeb MOHAMADI, *ICP - Université Stendhal*

Claude MONTACIE, *Université Paris VI*

José MORAIS, *Université Libre de Bruxelles*

Dominique MORIN, *C.N.E.T - Lannion*

Andrew MORRIS, *ICP - Université Stendhal*

Luc MORTIER, *Lernout & Hauspie Speech Products*

F. MOURIA, *Université de Nancy I*

Mohamad MRAYATI, *ISSAT - Damas*

Mohamed NAIT-LAHCEN, *LIMSI - BULL - Orsay*

Pierre NERZIC, *LLI-ENSSAT*

Noël NGUYEN-TRONG, *Université de Provence*

Yukihire NISHINUMA, *Université de Provence*

Régine André OBRECHT, *IRISA - CNRS - Rennes*

D. PASCAL, *C.N.E.T - Lannion*

Valérie PASDELOUP, *Université Libre de Bruxelles*

Jean-Marie PECATTE, *Université Paul Sabatier*

Guy PERENNOU, *Université Paul Sabatier*

Pascal PERRIER, *ICP - INP Grenoble*

Jean-Marie PIERREL, *Université de Nancy 1*

Michel PITERMANN, *Université Libre de Bruxelles*

Dolors POCH, *Université de Barcelone*

Monique RADEAU, *Université Libre de Bruxelles*

E. REYNIER, *ICP - INP Grenoble*

Annie RIALLAND, *Université de Paris III*

Gaël RICHARD, *LIMSI - CNRS - Orsay*

Véronique RISSOAN, *ICP - Université Stendhal*

Jordi ROBERT-RIBES, *ICP - INP Grenoble*

Xavier RODET, *Université de Paris VI*

L. ROMARY, *Université de Nancy I*

Marco SAERENS, *Lernout & Hauspie Speech Products*

Serge SANTI, *Université de Provence I*

Christophe SAVARIAUX, *ICP - INP Grenoble*

Jean Bernard SCHOENTGEN, *Université Libre de Bruxelles*

Jean-Luc SCHWARTZ, *ICP - INP Grenoble*

Célia SCULLY, *University of Leeds*

Christophe SEGEBARTH, *Université Libre de Bruxelles*

Jean-François SERIGNAT, *ICP - INP Grenoble*

Willy SERNICLAES, *Université Libre de Bruxelles*

Takahiko SHINMURA, *Université de Shiruoka*

K. SMAILI, *Université de Nancy I*

Alain SOQUET, *Université Libre de Bruxelles*

A. SOUBIGOU, *C.N.E.T - Lannion*

C. SPAGNOLETTI, *Université Libre de Bruxelles*

Thierry SPRIET, *Université d'Avignon*

Nelly SUAUDEAU, *IRISA - CNRS - Rennes*

Zakari TCHAGBALE, *Université Nationale de Côte d'Ivoire*

Jacques TERKEN, *Institut de Recherche en Perception - I.P.O.*

Bernard TESTON, *Université de Provence*

Jacqueline TIHONI, *Université Paul Sabatier*

Chantal TREPANIER, *Université de Montréal*

Jacqueline VAISSIERE, *Université de Paris III*

Nathalie VALLEE, *ICP - Université Stendhal*

B. VAN COILE, *Lernout & Hauspie Speech Products*

Joëlle VAN EIBERGEN, *ICP - Université Stendhal*

Ph. VERDIER, *ICP - INP Grenoble*

Nadine VIGOUROUX, *Université Paul Sabatier*

C. VLOEBERGHS, *Ecole Royale Militaire - Bruxelles*

Sophie WAUQUIER-GRAVELINE, *Université Paris VII*

Patrice WOODWARD, *ICP - Université Stendhal*

Mohamed YEOU, *Université de la Sorbonne Nouvelle*

J. ZEILIGER, *ICP - INP Grenoble*

Brigitte ZELLNER-BECHEL, *Université Paris VII*

Imad ZNAGUI, *Université de Paris III*

Mounir ZRIGUI, *IRSIT - Tunis*

Liste des Institutions

BULL
7 rue Ampère
F-91343 Massy Cedex
FRANCE

C.N.E.T - Lannion
LAA/TSS/RCP
Route de Trégastel - BP 40
F-22301 Lannion Cedex
FRANCE

CAP GEMINI INNOVATION
118 Rue Tocqueville
F-75017 Paris
FRANCE

CNRS - Paris
Labo de Psychologie expérimentale
Rue Serpente 28
F-75006 Paris
FRANCE

Ecole Royale Militaire
avenue de la Renaissance 30
B-1040 Bruxelles
BELGIQUE

ENST - Télécom Paris
Département Signal
46 Rue Barrault
F-75634 Paris Cédex 13
FRANCE

Faculté des Sciences de Luminy
G.I.A. - URA 816
163 avenue de Luminy - case 901
F-13288 Marseille cedex 9
FRANCE

ICP - INP Grenoble
ENSERG
46 Rue Félix Viallet
F-38031 Grenoble Cédex
FRANCE

ICP - Université Stendhal
Domaine universitaire BP 25X
F-38400 Grenoble Cedex
FRANCE

INRS-Télécommunications
3 Place du Commerce
Ile des Soeurs
Verdun
(Québec) H3H 1H6
CANADA

Institut de Recherche en Perception
IPO
P.O. Box 513
NL-5600 MB Eindhoven
NEDERLAND

Institut des Recherches Fondamentales de Technologie
Académie polonaise des Sciences
rue Swietokrzyska 21
00-049 Varsovie
POLOGNE

IRISA - CNRS - Rennes
Campus de Beaulieu
F-35042 Rennes Cédex
FRANCE

IRSIT - Tunis
Université de Tunis 1
2 rue Ibn Nadim - Cité Montplaisir
1002 Tunis
TUNISIE

ISSAT - Damas
BP 7028
Damas
SYRIE

Lernout & Hauspie Speech Products
Koning Albert I laan 64
B-1780 Wemmel
BELGIE

LIMSI - CNRS - Orsay
B.P. 30
F-91406 Orsay Cédex
FRANCE

LLI-ENSSAT
6 Rue de Kérampont
F-22305 Lannion cédex
FRANCE

Philips Research Laboratory
dissout fin 1991

U.S.T.H.B. - Alger
Institut d'Électronique
B.P. 32 EL-ALIA
Alger
ALGERIE

Université d'Avignon
Laboratoire d'Informatique
Faculté des Sciences
Rue Louis Pasteur 33
F-84000 Avignon
FRANCE

Université de Barcelone
S-08193 Bellaterra
Barcelona
ESPAGNE

Université de Caen
Dept. de Linguistique française
place de la Paix
F-14000 CAEN
FRANCE

Université de l'Etat de Mons
Faculté de Psycho. Pédagogie
Avenue du Champs de Mars
B-7000 Mons
BELGIQUE

Université de la Sorbonne Nouvelle
Institut Phonétique (URA 1027)
Rue des Bernardins 19
F-75005 Paris
FRANCE

Université de Montréal
Dép. de Linguistique
Case postale 6128 - succursale A
Montréal (Québec) H3C 3J7
CANADA

Université de Nancy 1
C.R.I.N.
BP 239
F-54506 Vandoeuvre Cédex
FRANCE

Université de Paris III
Institut d'Etudes Linguistiques et Phonétiques
CNRS-URA 1027
Rue des Bernardins 19
F-75005 Paris
FRANCE

Université de Paris VI
LAFORIA
CNRS URA 1095
4 Place Jussieu
F-75252 Paris Cedex 5
FRANCE

Université de Paris VII
Dépt. Recherches Linguistiques
Laboratoire de Phonétique
10 Rue Charles V
F-75004 Paris
FRANCE

Université de Paris-Sud
Laboratoire de Recherche en Informatique
U.A. 410 C.N.R.S.
Centre d'Orsay Bât. 490
F-91405 Orsay
FRANCE

Université de Provence
Institut de Phonétique
Labo. Parole et Langage
CNRS URA 261
29 Avenue Robert Schuman
F-13621 Aix-en-Provence
FRANCE

Université de Shiruoka
3-5-1 Johoku
Hamamatsu
JAPON

Université de Sofia
Dept. des Lettres classiques et modernes
Ivac Vojvoda 26
1124 Sofia
BULGARIE

Universität Trier
Sprach U. Literaturwissenschaften
Schneidershof
D55 TRIER
DEUTSCHLAND

Université La Sapienza
Département INFOCOM
Rome
ITALIE

Université Libre de Bruxelles
avenue F.D. Roosevelt 50
B-1050 BRUXELLES
BELGIQUE

Université Lumière Lyon 2
LAPHOLIA
86 rue Pasteur
F-69365 Lyon Cedex 07
FRANCE

Université Nationale de Côte d'Ivoire
Institut de Linguistique appliquée
08 BP 887
Abidjan 08
REPUBLIQUE DE COTE D'IVOIRE

Université Paul Sabatier
CERFIA
Route de Narbonne 118
F-31062 Toulouse Cédex
FRANCE

Université Pierre Mendès-Grenoble
Laboratoire de Psychologie Expérimentale
U.A. CNRS 665
BP 47 X
F-38040 Grenoble Cedex
FRANCE

Université Ramon Llull
Ecole Universitaire de Télécommunication "La Salle"
Passeig Bonanova 8
SP-08022 Barcelone (Catalogne)
ESPAGNE

University of Leeds
Department of Psychology
Leeds LS2 9JT
GRANDE-BRETAGNE