JEP 96

XXI^{es} JOURNÉES D'ÉTUDE SUR LA PAROLE

AVIGNON 10-14 JUIN 1996

XXIes JOURNÉES D'ÉTUDE SUR LA PAROLE

Organisées par

le Groupe Francophone de la Communication Parlée (GFCP)

de la Société Française d'Acoustique (SFA)
et de l'European Speech Communication Association (ESCA)

Avec le soutien
du GDR-PRC Communication Homme-Machine
et de l'Université d'Avignon et des Pays de Vaucluse

Centre d'Enseignement et de Recherche en Informatique

Technopôle Régional de l'Agroparc d'Avignon-Montfavet

10-14 juin 1996

XXIes JOURNÉES D'ÉTUDE SUR LA PAROLE

Organisées par

Le Groupe Francophone de la Communication Parlée (GFCP)

de la Société Française d'Acoustique (SFA)
et de l'European Speech Communication Association (ESCA)

avec le soutien

du GDR-PRC Communication Homme-Machine
et de l'Université d'Avignon et des Pays de Vaucluse

du 10 au 14 juin 1996

au Centre d'Enseignement et de Recherche en Informatique (CERI)
de l'Université d'Avignon et des Pays de Vaucluse (UAPV)

COMITÉ SCIENTIFIQUE

Régine ANDRÉ-OBRECHT (IRIT Toulouse)
Frédéric BIMBOT (ENST Paris)
Louis-Jean BOË (ICP Grenoble)
Jean-François BONNOT (IP Strasbourg)
Jean-Luc COCHARD (IDIAP Martigny)
Christophe D'ALESSANDRO (LIMSI Orsay)
Paul DELÉGLISE (LIUM Le Mans)
Pierre DUPONT (CNET Lannion)
Marc EL-BÈZE (LIA Avignon)
Daniel HIRST (LPL Aix-en-Provence)
Yves LAPRIE (CRIN Nancy)
Henri MÉLONI (LIA Avignon)
Claude MONTACIÉ (LAFORIA Paris)
Pascal NOCERA (LIA Avignon)
Pascal PERRIER (ICP Grenoble)
Jean SHOENTGEN (IP Bruxelles)
Bernard TESTON (LPL Aix-en-Provence)
Jacqueline VAISSIÈRE (IPP Paris)

COMITÉ D'ORGANISATION

Henri MÉLONI (Président)
Frédéric BÉCHET, Jean-François BONASTRE, Marc EL-BÈZE, Philippe GILLES
Stéphane IGOUNET, Philippe LANGLAIS, Alain SAMUEL, Thierry SPRIET

SECRÉTARIAT

Jocelyne GOURRET, Mireille ROZIER

TABLE DES MATIÈRES

PERCEPTION

Conférence Invité :

PRODUCTION

PHONÉTIQUE ET PHONOLOGIE

PROSODIE

PATHOLOGIE

ANALYSE ACOUSTIQUE

RECONNAISSANCE

Conférence Invité :

RECONNAISSANCE AUDIOVISUELLE

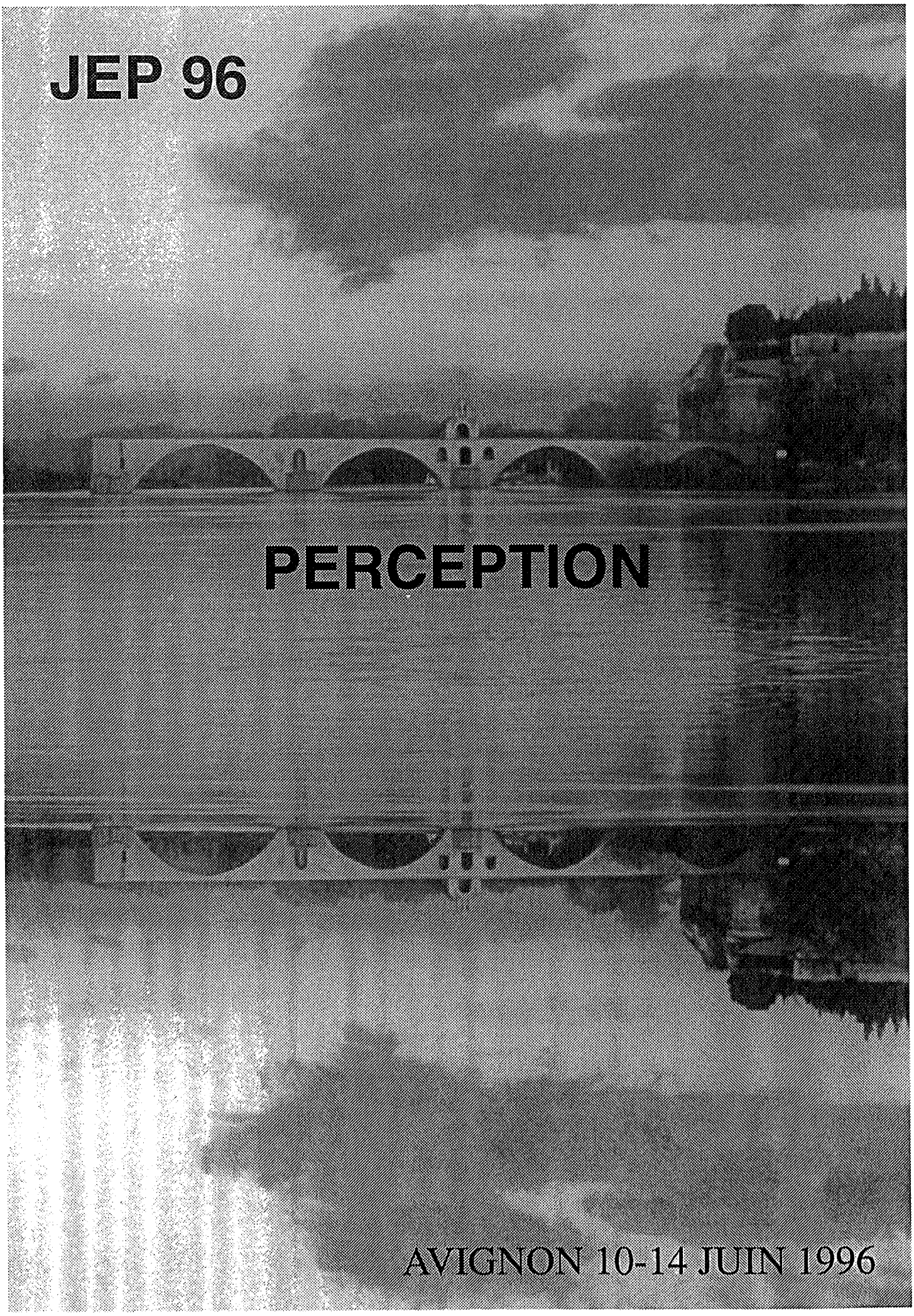
SYNTHÈSE

RECONNAISSANCE DU LOCUTEUR ET DE LA LANGUE

LEXIQUE ET DIALOGUE

RESSOURCES ET DÉMONSTRATIONS

JEP 96



PERCEPTION

AVIGNON 10-14 JUIN 1996

LA PERCEPTION DE LA PAROLE ET LA RECONNAISSANCE DES MOTS: APPORT D'UNE APPROCHE COMPUTATIONELLE

Uli Frauenfelder[1] et Alain Content[1,2]

1 Laboratoire de Psycholinguistique Expérimentale,
FAPSE - Université de Genève - 9 route de Drize - CH-1227 Carouge - SUISSE
Tel: (41) 22 705.97.410 - Fax: (41 22) 300.14.82 - e-mail: frauenfe@uni2a.unige.ch

2 Laboratoire de Psychologie Expérimentale, Université libre de Bruxelles - e-mail : acontent@ulb.ac.be

ABSTRACT

This tutorial paper presents some important issues raised in the area of spoken word recognition. It also introduces a novel psycholinguistic approach to addressing these issues that involves combining computational modelling, quantitative analyses of lexical databases, and experimentation. Finally, it is shown how this computational methodology can be used to investigate the two central problems of lexical activation and lexical segmentation.

1. INTRODUCTION

L'extrême efficience du locuteur humain pour comprendre le langage parlé dépend dans une large mesure de l'efficacité et de la rapidité des processus de reconnaissance des mots. Les psycholinguistes postulent l'existence d'un système de connaissances relatives aux mots —le lexique mental— qui sert à conserver différents types d'informations langagières (phonologiques, orthographiques, syntaxiques et sémantiques) en mémoire à long terme. L'association entre les informations sur la forme (phonologique et orthographique) des mots et sur la signification offre une solution naturelle au problème posé par le caractère arbitraire des relations entre forme et signification.

La description du système de traitement lexical suppose de déterminer l'organisation des entrées lexicales individuelles dans le lexique mental et leur structure interne, ainsi que la nature des codes utilisés et les mécanismes mis en oeuvre pour évoquer ces représentations et accéder aux différentes propriétés des mots.

L'objectif de cet article est de montrer comment une approche computationnelle peut contribuer à améliorer notre compréhension des mécanismes psychologiques opérant dans la perception de la parole et la reconnaissance des mots parlés. Dans la première partie, nous introduisons les questions principales qui sont examinées dans le domaine. Nous décrivons ensuite les difficultés méthodologiques associées à l'étude expérimentale du traitement lexical, et nous présentons différentes sources de données disponibles. Plus précisément, nous proposons une approche basée sur l'intégration entre trois techniques d'étude complémentaires: l'expérimentation humaine, l'analyse quantitative de bases de données lexicales informatisées, et l'utilisation de modèles de simulation. Enfin, nous présentons des exemples spécifiques de cette approche intégrée qui portent sur la question de l'activation et de la dé-activation des candidats lexicaux, et sur la question de la segmentation lexicale.

Trois questions centrales peuvent être envisagées à propos des processus de reconnaissance de la parole et des mots :

- 1) Comment la parole continue est-elle segmentée et classifiée en vue d'élaborer une représentation préalable à l'accès au lexique mental?
- 2) Quelle est la structure de la représentation pré-lexicale qui est élaborée à partir du signal acoustique pour accéder au lexique?
- 3) Quels sont les processus mis en jeu pour comparer la représentation construite à partir du signal et les représentations lexicales stockées en mémoire?

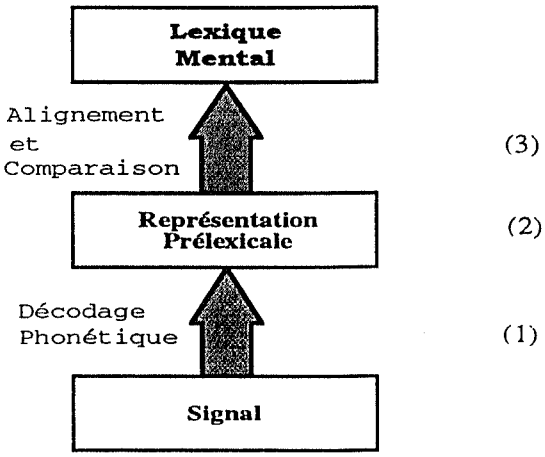


Figure 1. Représentation schématique des processus de reconnaissance des mots.

Ces trois questions peuvent être conceptualisées en considérant la caractérisation simplifiée des mécanismes de reconnaissance des mots présentée à la Figure 1.

Jusqu'à ces dernières années, la communauté scientifique a adopté une division des tâches assez nette entre les phonéticiens et les psycholinguistes. Les phonéticiens se sont plus intéressés à l'analyse du signal acoustique et à son codage phonétique, et se sont moins préoccupés des relations entre le décodage acoustico-phonétique et la reconnaissance des mots ou le traitement des phrases. Inversement, les psycholinguistes se sont plus souvent concentrés sur les étapes ultérieures du traitement, et ont parfois négligé les complexités du niveau acoustico-phonétique. Ce découpage artificiel laisse des failles entre les deux domaines d'étude et nous pensons qu'il est nécessaire de porter plus d'attention à l'interface entre le décodage acoustico-phonétique et le traitement lexical.

2. MÉTHODOLOGIES D'ÉTUDE

De plus en plus, les psycholinguistes développent une approche nouvelle de l'étude du traitement lexical qui conduit à combiner et intégrer les données expérimentales, l'analyse du comportement de modèles de simulation, et les descriptions quantitatives de bases de données lexicales. La figure 2 présente une description des interactions entre ces trois sources d'information.

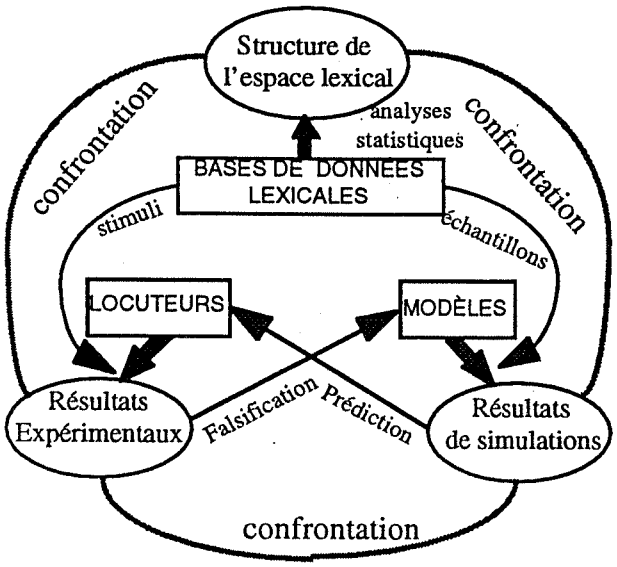


Figure 2. Schéma des relations envisagées entre les trois approches décrites.

2.1. Techniques expérimentales en temps réel

Ces techniques demandent aux locuteurs de fournir des réponses rapides à des stimuli de parole, et se basent sur la mesure des temps de réponse (TR) pour inférer les caractéristiques du traitement mis en jeu. Les techniques chronométriques constituent actuellement la méthodologie dominante de l'expérimentation psycholinguistique, particulièrement dans le champ de la reconnaissance des mots. Le recours à des tâches simples (détection, décision binaire, répétition immédiate) et la réduction de l'intervalle de temps séparant la présentation de la stimulation et la réponse exprimée par les participants permettent d'espérer que ces observations reflètent avec une certaine fidélité la nature et la durée des opérations mentales mises en oeuvre. Les techniques de détection de cibles et d'amorçage inter-modal sont particulièrement utilisées et seront décrites plus en détail dans les exemples qui suivent.

2.2. Modèles computationnels

Les modèles computationnels sont basés sur des programmes de traitement numérique ou symbolique qui simulent les mécanismes de traitement. Les techniques de modélisation fournissent un outil de choix pour prendre en compte des phénomènes d'une grande complexité. De nombreuses sciences, comme l'économie, la biologie ou la météorologie, ont évolué naturellement vers l'utilisation de techniques de simulation depuis que des ressources de calcul automatique d'une puissance suffisante sont disponibles. En psycholinguistique, les chercheurs ont également commencé à développer des systèmes de simulation du traitement du langage pour tenter de rendre compte de la complexité des phénomènes impliqués (Cf. Dijkstra & de Smedt, 1996).

On peut identifier plusieurs avantages importants à l'utilisation de modèles de simulation. D'une part, ils forcent le concepteur à préciser les détails des opérations, qui font souvent défaut dans les formulations verbales des théories. En outre, l'examen du comportement du système permet de vérifier si la théorie est complète et cohérente, et fournit un test préliminaire de sa plausibilité. Enfin, ces modèles permettent des prédictions quantitatives beaucoup plus précises, qui peuvent être directement comparées avec les données des expériences psycholinguistiques (Cf. Figure 2).

Dans le domaine de la reconnaissance des mots parlés, plusieurs modèles de simulation ont été proposés, et ont eu une influence considérable sur l'évolution des idées et les recherches entreprises. En particulier, TRACE (McClelland & Elman, 1986) et SHORTLIST (Norris, 1994) sont décrits de manière détaillée dans la suite.

2.3 Statistiques lexicales

L'impact des développements dans les technologies linguistiques est sensible dans le domaine de la psycholinguistique, notamment

à travers l'utilisation accrue de bases de données lexicales. Celles-ci deviennent un outil indispensable de la recherche expérimentale. Les bases de données lexicales contiennent divers types d'informations (informations phonologiques, orthographiques, morphologiques, syntaxiques et sémantiques, comptages de fréquence). Des bases de données existent actuellement pour de nombreuses langues. CELEX (Celex, 1993) inclut l'anglais, l'allemand et le néerlandais, et BRULEX (Content, Mousty & Radeau, 1990) porte sur le Français.

L'analyse statistique de bases de données lexicales permet de produire des descriptions quantitatives des propriétés structurales des langues à différents niveaux, ainsi que la réalisation de comparaisons des caractéristiques de langues différentes (Cf. par exemple Goldman, Content & Frauenfelder, ce volume). De plus, les bases de données lexicales remplissent deux autres fonctions plus pragmatiques, mais néanmoins cruciales pour l'expérimentation et la simulation : la sélection de stimuli linguistiques utilisés dans l'expérimentation avec tous les contrôles appropriés des variables parasites potentielles (Cutler, 1980), et la génération de sous-lexiques représentatifs qui peuvent être incorporés dans les études de simulation.

Ces trois sources d'information peuvent donc être combinées et comparées de diverses manières. A titre d'illustration de leur complémentarité, nous présentons deux exemples de recherches récentes. Le premier porte sur le problème de la segmentation lexicale et le second sur les mécanismes d'activation et d'élimination des candidats lexicaux au cours de la présentation des mots parlés.

3. LA RECONNAISSANCE DES MOTS

Il existe à l'heure actuelle un accord assez large sur l'idée que la reconnaissance des mots suppose l'activation d'un ensemble de candidats potentiels, et la sélection du mot cible dans cet ensemble. Cette conception découle assez naturellement de la nature continue et séquentielle du signal de parole. La plupart du temps, les locuteurs ne disposent que d'une information sensorielle partielle relative au stimulus, et cette information est souvent insuffisante pour identifier le mot. On peut donc supposer que le locuteur génère de manière immédiate et continue des hypothèses sur base de cette information partielle.

Deux questions essentielles qui apparaissent dans ce cadre sont examinées dans les sections suivantes. La première concerne la détermination des candidats activés —éventuellement de manière très momentanée— durant le processus de reconnaissance. La seconde concerne la sélection finale de l'entrée lexicale appropriée et le mécanisme de rejet des candidats inappropriés.

3.1. Alignement et segmentation lexicale

Pour identifier les mots effectivement entendus, le locuteur est confronté à la tâche de trouver l'alignement correct entre les représentations prélexicales dérivées de l'information sensorielle et les représentations lexicales. En d'autres mots, le locuteur doit déterminer quelle partie de l'information sensorielle doit être comparée avec les représentations lexicales. Le fait que la parole est continue et que les frontières des mots ne sont pas marquées explicitement de manière systématique (Cole & Jakimik, 1980) rend le mécanisme d'alignement plus ardu. Trois types de propositions existent dans la littérature psycholinguistique sur cette question.

Selon l'hypothèse d'alignement positionnel, seules les entrées lexicales qui sont alignées sur une position spécifique de l'information sensorielle sont prises en considération. Une approche consiste à considérer que seules les hypothèses lexicales qui correspondent aux débuts de mots sont activées. Ainsi, le modèle de la Cohorte (Marslen-Wilson, 1984) est basé sur le principe d'alignement des débuts de mots. En effet, selon cette théorie, les parties du signal correspondant à des débuts de mots sont les seules à déterminer l'évocation de candidats lexicaux. Les candidats qui ne correspondent pas ne sont jamais activés. Cette hypothèse présuppose un moyen de localiser de façon robuste les frontières des mots dans l'information sensorielle.

L'espace de recherche lexicale peut aussi être réduit aux mots qui correspondent à certaines parties du signal, sans tenir compte de la position de ces parties. Un exemple de ce principe d'alignement par repères peut être trouvé dans le modèle proposé par Grosjean & Gee (1987), dans lequel les syllabes accentuées, perceptivement plus saillantes, constituent des points d'alignement. Ainsi, toutes les entrées lexicales comportant une syllabe accentuée particulière, quelle que soit sa position dans le mot, seront des candidats potentiels pour la reconnaissance. Les notions d'alignement positionnel et d'alignement de repères peuvent être combinées. C'est le cas pour l'hypothèse d'une Stratégie de segmentation métrique avancée par Cutler & Norris (1988) spécifiquement pour l'anglais. Selon cette hypothèse, on suppose également

que le locuteur utilise des syllabes fortes comme repères, mais en outre, on postule que ces syllabes accentuées correspondent à des débuts de mots, de sorte que les seuls candidats lexicaux évoqués sont ceux dont le début est aligné sur la syllabe repérée et y correspond. (Pour une discussion des indices prosodiques utilisés dans la segmentation du français, cf. Banel et Bacri (1994) et Meunier, Frauenfelder & Content, ce volume).

Enfin, une approche plus radicale, l'alignement exhaustif, envisage la possibilité que chaque unité extraite de l'information sensorielle soit comparée à l'ensemble des entrées lexicales. Si chaque correspondance entre un segment du signal et une représentation lexicale détermine l'activation du candidat, il est évident qu'un nombre extrêmement grand de candidats (inappropriés) seront activés. Le modèle d'activation interactive TRACE adopte ce principe dans la mesure où tous les mots du lexique sont constamment en lice pour la reconnaissance, leur état d'activation augmentant et diminuant de manière continue à chaque cycle de traitement et en fonction du signal reçu et des compétiteurs activés. On verra plus loin qu'en pratique, tous les candidats ne sont pas forcement activés.

Ces différentes hypothèses conduisent à des prédictions en partie différentes sur la nature de l'ensemble des mots qui sont activés par une stimulation de parole. L'examen de ces questions et le choix entre différentes possibilités peut tirer parti de l'approche triple que nous avons décrite précédemment.

Des analyses statistiques sur BRULEX nous permettent de déterminer globalement combien d'entrées lexicales seraient activées, sur base d'hypothèses particulières sur l'alignement et la comparaison. Le recours aux simulations, dans le cas présent, avec deux modèles, TRACE et SHORTLIST, fournit une description quantitative précise de l'augmentation de l'activation et de sa décroissance au cours du temps pour chaque entrée lexicale. Ces résultats de simulation peuvent être directement comparés avec des données empiriques obtenues sur des locuteurs humains.

3.1.1. Statistiques lexicales

En principe, les modèles supposant un alignement exhaustif prédisent que toutes les représentations lexicales qui correspondent à une partie de la chaîne d'entrée seront activées, même lorsqu'elles correspondent à un mot enchâssé dans un autre mot plus long ou à un chevauchement entre deux mots différents. L'utilisation de bases de données lexicales permet d'estimer l'extension de ces phénomènes d'enchâssement, tandis que les chevauchements peuvent être calculés à partir de corpus de textes continus.

La figure 3 présente les résultats principaux d'une analyse de l'enchâssement lexical pour le français, à partir de la base de données BRULEX. L'analyse est basée sur les mots de longueur égale ou supérieure à deux phonèmes qui se trouvent enchâssés dans des mots de deux syllabes ou plus. Il faut noter que le nombre de mots enchâssés est augmenté dans ces calculs par la présence dans BRULEX de formes morphologiquement liées ("nation" se trouve donc dans "national"), mais qu'il est par ailleurs diminué du fait que les formes fléchies (verbes conjugués, féminins, etc.) ne sont pas toutes incluses dans la base de données.

Les résultats montrent un très haut degré d'enchâssement dans le lexique français. Environ la moitié des mots contiennent au moins cinq mots enchâssés et plus de 99% des mots contiennent au moins un mot enchâssé.

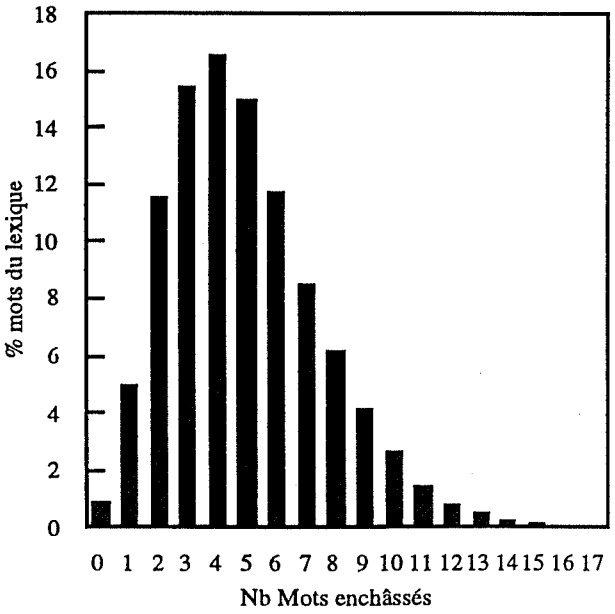


Figure 3. Pourcentage de mots qui contiennent un nombre donné de mots enchâssés.

Le nombre de candidats potentiels est bien sûr encore beaucoup plus élevé pour la parole continue, puisque les chevauchements viennent s'ajouter aux enchâssements. Harrington & Johnstone (1987) ont calculé l'extension de l'ensemble des candidats lexicaux potentiels pour un ensemble de 115 phrases anglaises transcrites phonémiquement. Les phrases variaient en longueur de quatre à six mots. D'après l'analyse que nous avons faite de leurs résultats, plus de 45% des phrases contenaient plus de 100 candidats, (soit environ 20 fois le nombre réel de mots).

Ces statistiques indiquent l'importance massive des phénomènes d'enchâssement et de chevauchement. Elles démontrent que les mécanismes de reconnaissance des mots devraient gérer un grand nombre d'hypothèses si tous les candidats enchâssés étaient évoqués, et indiquent donc la nécessité d'envisager comment les modèles du traitement lexical peuvent contraindre l'étendue de l'ensemble des candidats activés.

3.1.2. Modélisation

Les modèles computationnels permettent d'aller au-delà de ce type d'analyse quantitative en mettant en oeuvre des hypothèses détaillées sur les mécanismes de traitement. L'utilisation de simulations permet ensuite d'évaluer les conséquences dynamiques complexes de ces hypothèses sur les performances de reconnaissance.

Le modèle TRACE est basé sur la notion d'activation interactive. Il est composé de trois niveaux d'unités correspondant à des Traits distinctifs, des Phonèmes et des Mots. Chaque unité représente une hypothèse relative à la stimulation fournie. Les trois niveaux sont organisés de façon hiérarchique (Cf Figure 4).

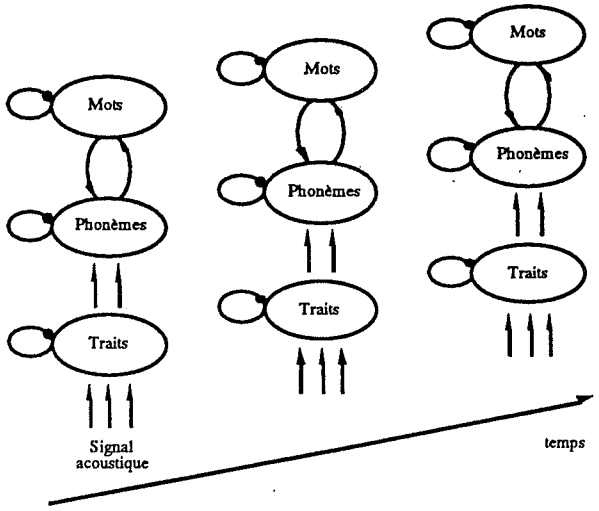


Figure 4. Le modèle TRACE.

La Figure 4 illustre de manière simplifiée les trois niveaux de représentation et la configuration de connectivité proposée. Elle indique aussi comment TRACE gère la dimension temporelle par réduplication du réseau pour les instants successifs.

Le système comporte des connexions excitatrices montantes et descendantes (traits-phonèmes, phonèmes-mots, et mots-phonèmes), et des connexions inhibitrices entre unités d'un même niveau (traits-traits, phonèmes-phonèmes, mots-mots). L'information sensorielle est simulée par l'excitation de certaines unités Traits, qui transmettent ensuite des signaux excitateurs aux unités Phonèmes. Les unités Phonèmes sont activées en fonction du degré de correspondance avec l'ensemble des traits distinctifs activés, de telle sorte que plusieurs unités phonèmes sont activées pour un input donné. Au fur et à mesure que l'activation des unités phonèmes croît, ceux-ci contribuent à augmenter le niveau d'activation des mots qui les contiennent. A leur tour, dès que les unités mots atteignent un certain niveau d'activation, elles commencent à s'inhiber entre elles (du fait des connexions inhibitrices entre les unités mots), et contribuent à augmenter, via les connexions descendantes, le niveau d'activation des phonèmes compatibles.

Le modèle SHORTLIST (Norris, 1994) constitue une tentative d'amélioration de certaines caractéristiques insatisfaisantes de TRACE, en particulier, l'implausibilité de son architecture en ce qui concerne la représentation de la dimension temporelle. SHORTLIST comporte deux phases de traitement distinctes, pour chaque segment successif de l'input. Durant la première étape, un ensemble limité de candidats est sélectionné (la liste de candidats). Tout mot, quel que soit son alignement avec l'input, peut être incorporé pour autant qu'il atteigne un certain degré de correspondance pré-établi (qui constitue un des paramètres du modèle) avec l'input. La comparaison entre l'information sensorielle et les représentations lexicales tient compte la fois des segments qui correspondent (qui contribuent à augmenter le score associé à chaque mot) et de ceux qui ne correspondent pas (sous forme de pénalisation). Durant la seconde étape, les candidats les plus appropriés —quel que soit leur alignement— sont soumis à un mécanisme de compétition basé sur le principe d'inhibition réciproque, comme dans TRACE, de sorte que les candidats les plus appropriés sur base de l'information sensorielle inhibent plus leurs compétiteurs. Cependant, à la différence de TRACE, dans lequel l'ensemble des mots du lexique sont pré-cablés dans le réseau, SHORTLIST établit une limite supérieure (variant entre 3 et 30 candidats) sur le nombre de mots pris en considération pour l'étape de compétition. Au cycle suivant, l'établissement de la liste tient compte à la fois de l'activation des candidats résultant de la compétition antérieure, et du degré d'appariement mis à jour en fonction du nouveau segment

Plus récemment, le modèle SHORTLIST a été partiellement modifié (Norris, McQueen, & Cutler, 1995). Une des différences, qui est pertinente pour la discussion qui suit, est que

la liste établie à chaque cycle ne tient plus compte du résultat du cycle précédent. Cette modification évite que le modèle s'engage trop rapidement en faveur de candidats qui seraient hautement activés sur base des premiers segments de la stimulation, et permet à des mots nouveaux d'entrer plus tardivement dans la liste. De ce fait, et à la différence de TRACE, SHORTLIST rend possible, dans une certaine mesure, l'activation de candidats enchâssés en position non-initiale.

3.1.3. Modélisation et simulations

Frauenfelder & Peeters (1990) ont réalisé une série de simulations avec TRACE pour examiner en détail l'activation des candidats selon leur alignement. La figure 5 illustre les résultats obtenus pour une comparaison de l'évolution de l'activation pour des candidats enchâssés au début ou à la fin dans des séquences qui correspondaient ou non à un mot.

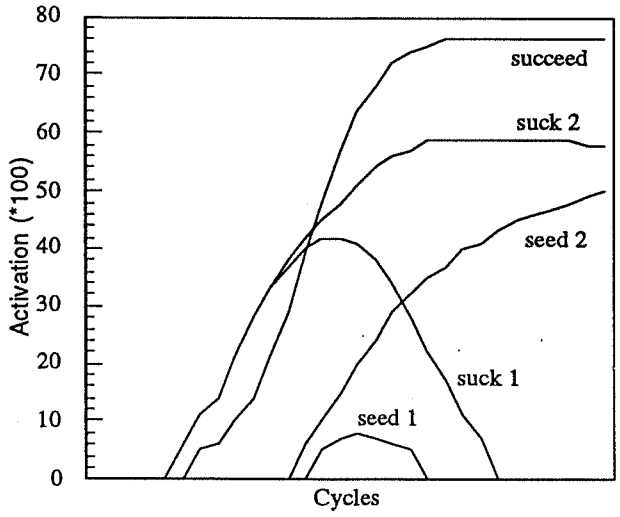


Figure 5. Courbes d'activation pour des mots enchâssés au début ("suck") et à la fin ("seed") dans une séquence ("succeed") selon que le mot correspondant à la séquence entière fait partie du lexique (1) ou qu'il en a été éliminé (2).

La Figure 5 montre de grandes différences dans les courbes d'activation correspondant aux mots enchâssés selon que la séquence constitue un mot ou non. Dans ce dernier cas, les deux candidats ("suck" et "seed") sont clairement activés, tandis que lorsque la séquence forme un mot, le candidat enchâssé au début est rapidement inhibé par le mot correspondant à la séquence entière. La domination du mot long est telle que l'activation du second mot enchâssé ne dépasse que brièvement le niveau de repos. Les différences d'activation entre les compétiteurs enchâssés, en fonction notamment de leur position dans la séquence illustrent le rôle crucial de l'inhibition entre les unités mots pour limiter le nombre de candidats évoqués.

Des simulations analogues ont été menées avec SHORTLIST, dans sa version récente. Comme on l'a vu, du fait que la liste de candidats est recalculée à chaque cycle, les candidats enchâssés en position non-initiale ont en principe plus de chances d'émerger malgré la présence d'une hypothèse dominante.

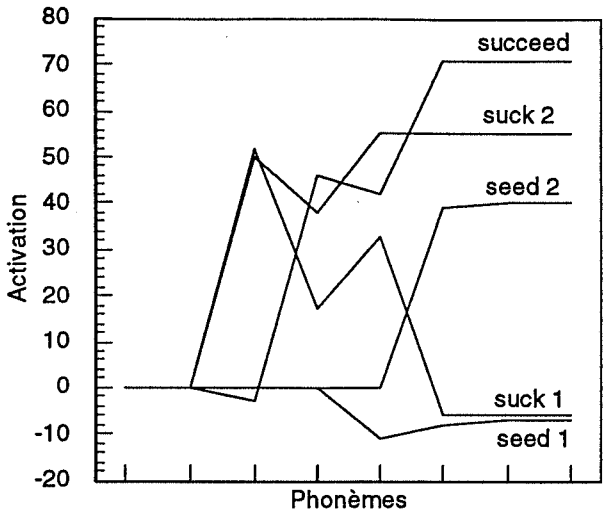


Figure 6. Courbes d'activation pour des mots enchâssés avec SHORTLIST.

Cependant, contrairement à cette attente, les simulations montrent des résultats analogues à ceux observés pour TRACE (Cf. figure 6). Pour les deux modèles, il semble donc que, lorsque la séquence entière forme un mot, les mots enchâssés ne soient activés que pour la position initiale. Par contre, lorsqu'il s'agit de séquences non mots, les deux mots enchâssés sont activés. Cette dernière prédiction a été testée expérimentalement.

3.1.4. Expérimentation

Une caractéristique attractive des modèles computationnels est la possibilité de générer des prédictions quantitatives précises sur l'activation des candidats et la reconnaissance. Frauenfelder & Henstra (1988) ont essayé d'établir si les mots enchâssés en position initiale ou finale sont effectivement activés durant le traitement. L'expérience exploitait une situation de détection de phonèmes, en néerlandais. Les participants devaient détecter, le plus rapidement possible un phonème qui leur était spécifié au préalable. Une estimation de l'activation lexicale était obtenue en comparant la différence des temps de détection dans des mots (ex., cible /p/: "mop", balai) et des pseudo-mots monosyllabiques (ex. "nop"). L'expérience comparait trois situations. la ligne de base ("mop" vs. "nop"), l'enchâssement initial ("mopel", vs. "nopel") et

l'enchâssement final ("temop" vs. "tenop"). Dans la mesure où les propriétés acoustiques et le contexte phonétique local des phonèmes cibles est stable, un avantage éventuel pour les mots peut être interprété comme un indice de l'activation de l'entrée lexicale correspondante. La figure 7 résume les résultats expérimentaux.

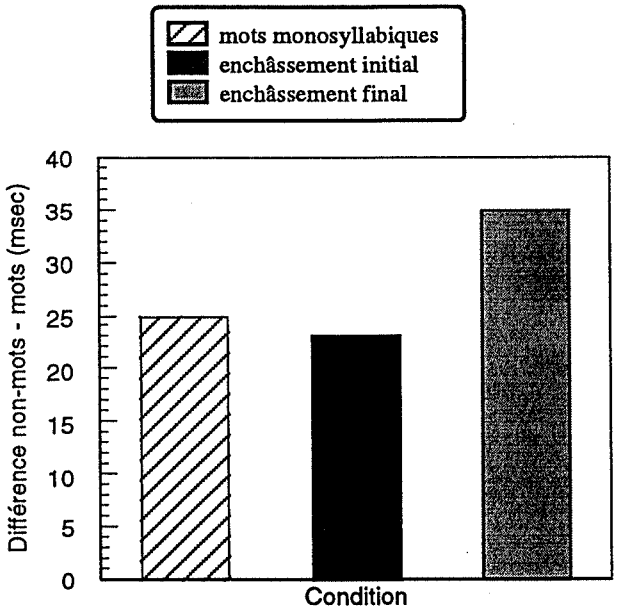


Figure 7. Différences entre les temps de détection de phonèmes (msec.) pour les mots et les non-mots, dans les trois conditions: mots monosyllabiques, enchâssement initial, et enchâssement final.

Les différences significatives observées dans les trois conditions suggèrent l'activation des mots enchâssés dans une séquence non-mot. Les résultats des simulations avec les deux modèles considérés, tout en différant dans le détail, sont compatibles avec ces données puisqu'ils montrent également une activation des candidats enchâssés dans la séquence.

En ce qui concerne l'activation de mots enchâssés dans des séquences qui constituent des mots, Shillcock (1990) a obtenus des résultats qui suggèrent que les candidats lexicaux enchâssés en position non-initiale seraient activés durant le traitement. Il a utilisé une procédure d'amorçage sémantique inter-modal, et a observé une facilitation significative du traitement d'un associé sémantique (ex. "rib", "côte de boeuf") lié à la deuxième syllabe d'un mot bisyllabique ("trombone", avec "bone", "os"). S'il était confirmé, ce résultat poserait des problèmes pour les deux modèles. Il est intéressant cependant de signaler que les stimuli expérimentaux utilisés par Shillcock comportaient en général une seconde syllabe accentuée. Il est donc possible qu'une simulation avec une version de SHORTLIST prenant en compte la stratégie de segmentation métrique rende compte de ce type de résultat.

En conclusion, nous avons étudié ici le problème de l'alignement lexical et de l'activation à travers trois approches différentes. L'accent a été mis sur l'alignement exhaustif, tel qu'il est mis en oeuvre dans les modèles d'activation interactive, comme TRACE ou SHORTLIST. Les résultats préliminaires, qui sont cohérents avec les deux modèles, suggèrent que l'auditeur génère constamment plusieurs hypothèses lexicales à partir de l'information sensorielle partielle et de segmentations potentielles. Si cette hypothèse est exacte, il sera nécessaire de comprendre comment l'auditeur, à partir d'un grand nombre de candidats possibles, arrive à identifier sans ambiguïté les mots dans la parole continue.

3.2. Croissance et décroissance de l'activation

Tout modèle de reconnaissance des mots est confronté à un dilemme: la probabilité que le mot correct soit inclus dans l'ensemble des candidats activés augmente avec le nombre de candidats, mais par contre, la probabilité que la sélection converge vers la solution correcte diminue avec le nombre de candidats. Si le système est conçu de façon à maximiser le nombre de candidats possibles, comme dans le cas de l'alignement exhaustif, une méthode efficace doit alors être appliquée pour éliminer les candidats inappropriés.

SHORTLIST et TRACE font appel à des solutions légèrement différentes pour réduire le nombre de candidats activés. Dans TRACE, la réduction de cet ensemble a lieu exclusivement par un mécanisme d'inhibition latérale. Cette inhibition entre les compétiteurs lexicaux permet aux candidats les plus forts de dominer et d'éliminer les candidats plus faibles. Selon le modèle SHORTLIST, la réduction de l'ensemble des candidats dépend à la fois de l'inhibition latérale et des pénalisations associées à la déviation entre les candidats et l'information sensorielle.

3.2.1. Modélisation et simulations

Dans les simulations suivantes, nous avons comparé les prédictions que font les deux modèles à propos de l'influence d'une information sensorielle divergente sur l'activation lexicale. Nous avons notamment examiné l'effet d'une substitution phonémique (ex. /n/ dans "vocabunaire") sur le niveau d'activation du mot "vocabulaire". Les

résultats de la simulation avec les deux modèle sont présentés à la Figure 8.

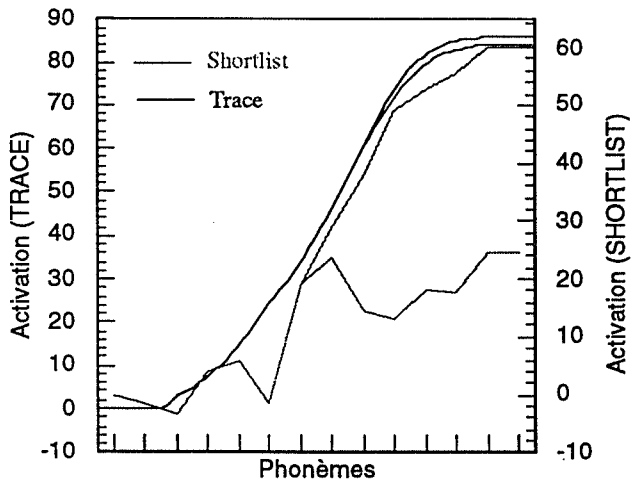


Figure 8. Courbes d'activation obtenues avec TRACE et SHORTLIST pour les mots cibles selon que la séquence présentée correspond exactement au mot ("vocabulaire") ou comporte un phonème déviant ("vocabunaire").

Pour TRACE, aucune différence n'apparaît entre les deux stimulations, ce qui montre clairement qu'une déviation tardive ne provoque pas de décrue de l'activation. Cela s'explique par le fait qu'au moment où l'information déviante parvient au système, la cible est déjà le candidat lexical le plus activé: il n'existe aucun autre mot qui corresponde à l'entrée erronée, et qui puisse inhiber l'activation de la cible lexicale.

En ce qui concerne SHORTLIST, les résultats montrent au contraire qu'une information sensorielle déviante a une forte capacité inhibitrice. En effet, dès que le phonème substitué parvient au système, le niveau d'activation du mot cible chute de manière importante. Il est intéressant de remarquer que ce niveau se rétablira peu de temps après à sa valeur initiale.

Ces simulations montrent que des modèles basés sur des mécanismes de sélection différents font des prédictions divergentes à propos de la nature des facteurs susceptibles d'affecter le processus de sélection et le déroulement temporel de la reconnaissance des mots.

Pour des modèles comme SHORTLIST qui admettent qu'une information sensorielle divergente a un effet inhibiteur, n'importe quelle information sensorielle déviante abaisse directement le niveau d'activation des candidats qui ne correspondent pas. Par contre, pour des modèles comme TRACE, l'information déviante n'affecte le processus de reconnaissance des mots que s'il y a d'autres candidats lexicaux qui correspondent à cette information, et sont activés par la nouvelle entrée. De tels candidats vont inhiber le mot cible. Si par contre il n'y a pas de compétiteurs, alors l'information erronée n'aura aucun effet.

3.2.2. Expérimentation

Pour mettre à l'épreuve les prédictions divergentes des deux modèles, nous avons utilisé à nouveau la tâche de détection de phonèmes (Frauenfelder, Content, Scholten, 1996). L'expérience visait à évaluer si une déviation tardive provoquait une diminution sensible de l'activation du mot cible. Nous avons donc comparé les temps de détection d'un phonème situé après le phonème substitué (ex. /R/ dans *vocabunaire*) et ceux obtenus pour le même phonème dans le mot intact (/R/ dans *vocabulaire*), par rapport à des non-mots comparables pour le contexte local (respectivement *satodunaire* et *satodulaire*).

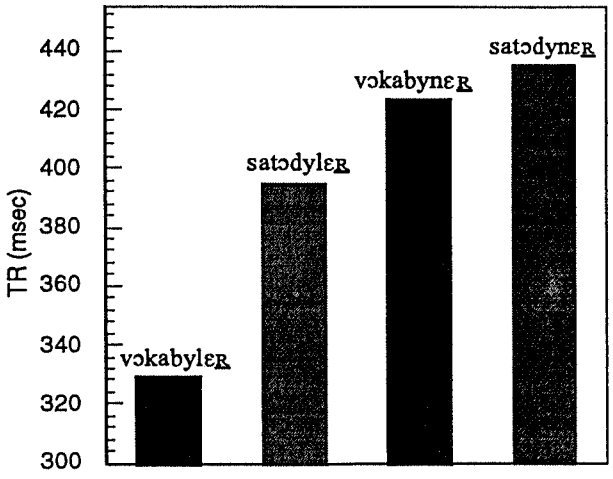


Figure 9: Temps moyens de détection de phonèmes dans les 4 conditions expérimentales.

Les résultats principaux apparaissent à la Figure 9. L'expérience a montré une différence significative entre les mots et les déviations minimales (*vocabulaire* vs *vocabunaire*), tandis que la performance pour les déviations minimales ne différait pas significativement de celle observée pour les non-mots contrôles. Ce résultat indique que l'activation du mot cible a décru rapidement suite à la réception du phonème divergent. Comme on l'a vu, SHORTLIST prédit une telle diminution, contrairement à TRACE.

4. CONCLUSIONS

Nous avons signalé la séparation existant entre l'étude du décodage acoustico-phonétique et l'étude des mécanismes d'accès au lexique. De plus en plus, les psycholinguistes tentent de prendre en considération les deux problématiques simultanément. La place manque pour

analyser en détail ces travaux, mais on a noté que le problème de la segmentation lexicale se pose dans des termes différents si l'on suppose que des indices métriques et prosodiques sont disponibles et font partie intégrante de l'information utilisée par le locuteur pour accéder au lexique. Enfin, du point de vue méthodologique, nous avons insisté sur l'intérêt d'une meilleure intégration entre le travail de description des caractéristiques des lexiques, le développement de modèles de simulation, et les études expérimentales. Nous pensons que c'est seulement à travers la combinaison de ces différentes sources de données que l'étude du traitement lexical humain pourra progresser.

REMERCIEMENTS

Cette recherche a été financée par le F.N.R.S. (Projet 11-39553.93). Nous remercions J.P. Goldman D. Norris, C. Meunier et M. Scholten pour leur aide dans différents aspects de ce travail.

BIBLIOGRAPHIE

Banel, M-H., & Bacri, N. (1994). On metrical patterns and lexical parsing in French. *Speech Communication*, 15,115-126.

Celex (1993). Centre for lexical information. Max-Planck-Institute for Psycholinguistics, Nijmegen, NL.

Cole, R. A. & Jakimik, J. (1980). A model of speech perception. In R.A. Cole (Ed.), *Perception and production of fluent speech.* Hillsdale, N.J.: Lawrence Erlbaum Associates.

Content, A, Mousty, P. & Radeau, M. (1990). Brulex: Une base de données lexicales informatisée pour le français écrit et parlé. *Année Psychologique*, 90, 551-556.

Cutler, A. (1981). Making up materials is a confounded nuisance, or: Will we be able to run any psycholinguistic experiments at all in 1990? *Cognition* ,10, 65-70.

Cutler, A. & Norris, D. (1988). The role of the strong syllables in segmentation for lexical access. *Journal of Experimental Psychology: Human Perception and Performance*, 14, 1, 113-121.

Dijkstra, A. & de Smedt, K. (1996). *Computational Psycholinguistics.* Prentice Hall .

Frauenfelder, U. H., Content, A., Scholten, M. (1996). Lexical Activation and Deactivation in Spoken Word Recognition (en préparation).

Frauenfelder, U. H. & Henstra, J. (1988). Activation and deactivation of phonological representations. *Proceedings of the 4th International Phonology Congress* , Krems, Austria.

Frauenfelder, U. H. & Peeters, G. (1990). On lexical segmentation in TRACE: An exercise in simulation. In G. Altmann (Ed.), *Cognitive models of speech processing: Psycholinguistic and Computational Perspectives* . Cambridge, MA: MIT Press.

Goldman, J.-P., Content, A. & Frauenfelder, U.H. (ce volume).

Grosjean, F. & Gee, J.P. (1987). Prosodic structure in spoken word recognition. *Cognition* , 25, 1-2, 157-187.

Harrington, J. & Johnstone, A.M. (1987). The effects of equivalence classes on parsing phonemes into words in continuous speech recognition. *Computer, Speech and Language*,22, 273-288.

Marslen-Wilson, W. D. (1984). Function and process in spoken word recognition. In H. Bouma & D. G. Bouwhuis (Eds.), *Attention and Performance X: Control of Language Processes* . Hillsdale, N. J.: Lawrence Erlbaum Associates. 125-149.

McClelland, J. L. & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, 18, 1-86.

Meunier, C. Frauenfelder, U.H., & Content, A. (ce volume).

Norris, D. (1994). SHORTLIST: A connectionist model of continuous speech recognition. *Cognition*, 52,189-234.

Norris, D., McQueen, J. M., & Cutler, A. (1995). Competition and segmentation in spoken word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 21, 1209-1228.

Shillcock, R. C. (1990). Lexical hypotheses in continuous speech. In G. Altmann (Ed.) *Cognitive models of speech processing: Psycholinguistic and computational perspectives.* Cambridge, MA: MIT Press.

AMORÇAGE DE REPETITION ET COMPETITIONS LEXICALES

Frédéric ISEL et Nicole BACRI

Université René Descartes, Laboratoire de Psychologie Expérimentale, URA 316 CNRS,
28 rue Serpente, F-75006 Paris
Tel: (1) 40 51 98 65 - Fax: (1) 40 51 70 85 - e-mail: isel@ext.jussieu.fr

ABSTRACT

Access to short words embedded in trisyllabic words has been studied in two cross-modal repetition priming experiments. A *long-word advantage* was found, as usual, when the visual target appeared at the offset of the auditory prime. When a short interval of silence was introduced between the embedded components, results show a decrease of *long-word advantage* and a strong inhibitory effect for the two shorter words. Data demonstrate the inhibitory effect of competition between multiple lexical hypotheses during word recognition.

Keys words: speech perception, segmentation, word recognition, cross-modal priming.

1. INTRODUCTION

Afin de comprendre la parole, les auditeurs doivent segmenter le signal acoustique continu en entités lexicales discrètes perçues de façon claire et immédiate. Cette tâche n'est pas triviale puisque les indices acoustiques de frontières de mots sont relativement peu fiables (Klatt, 1980). Dans la présente étude, le problème de la segmentation lexicale sera abordé en traitant de l'ambiguité lexicale inhérente à certains mots trisyllabiques français dans lesquels sont enchâssés un mot monosyllabique en première position et un mot bisyllabique en seconde position. Ces mots constituent un matériel de choix pour l'étude de la segmentation puisqu'en effet, un support phonétique analogue peut conduire à l'accès à différents mots (Frauenfelder et Peeters, 1990). Le mot « pantalon » illustre précisément ce point: en effet, confrontés à cette séquence de phonèmes, les auditeurs percoivent-ils uniquement le mot trisyllabique porteur /pantalon/ ou accèdent-ils, même de façon transitoire, aux deux composants enchâssés, en l'occurrence /pan/ et /talon/? Le français présente 85 % de mots plurisyllabiques, ce qui suggère que l'enchâssement lexical est probablement très fréquent. Ce type d'ambiguités lexicales offre l'opportunité de suivre le jeu des compétitions entre les différentes hypothèses lexicales (Shillcock, 1990). Une solution pour résoudre cette ambiguité peut être l'utilisation d'une routine spécifique de segmentation (Banel et Bacri, 1994). Le modèle TRACE (McClelland et Elman, 1986) propose une autre solution: la segmentation et la reconnaissance d'un mot sont le produit d'excitations inter-niveaux et d'inhibitions intra-niveaux et émergent de processus d'activation interactifs. La reconnaissance des mots se fait de manière interactive grâce à un flux continu et bidirectionnel d'informations entre le niveau de traitement lexical et le niveau de traitement acoustico-phonétique. L'accès au lexique est multiple et non séquentiel et la notion de « recherche lexicale en parallèle » est fondamentale.

L'objectif de cette recherche est de déterminer si des auditeurs confrontés à des séquences plurisyllabiques contenant des mots enchâssés segmentent ou non les mots porteurs en leurs deux composants. En référence à une architecture connexionniste, au niveau lexical, c'est le mot le plus long qui recevra la plus grande activation et de ce fait inhibera les mots enchâssés qui le constituent. Le composant monosyllabique occupant la position initiale dans le mot porteur est peut-être activé mais d'une façon transitoire. Cette activation du trisyllabe porteur devrait être mise en évidence par un effet d'amorçage lors d'une tâche d'amorçage inter-modal de répétition.

2: EXPERIENCE 1

L'objectif de cette expérience est d'évaluer le jeu des compétitions entre différentes

hypothèses lexicales lors du traitement de séquences sonores ambiguës. La technique de l'amorçage inter-modal de répétition a été utilisée. Selon Marslen-Wilson (1990), elle permet d'étudier les effets transitoires de l'activation de multiples hypothèses lexicales suscitées en différents points du signal. Les sujets entendent une amorce auditive suivie immédiatement après d'une cible visuelle et ils effectuent une tâche de décision lexicale sur la cible visuelle (décider si la séquence de lettres lue est ou non un mot). L'activation résiduelle de la représentation lexicale correspondant à l'amorce facilitera le traitement d'une cible visuelle reliée à l'amorce. Cette facilitation sera mise en évidence par des temps de réponse plus rapides pour les cibles visuelles reliées aux amorces que pour celles ne l'étant pas.

2.1. Méthode

2.1.1. Condition expérimentale

Sujets: 30 étudiants de langue maternelle française ne présentant aucun trouble de l'audition et de la vision

Matériel

Les amorces: 21 mots trisyllabiques de fréquence basse contenant chacun un mot monosyllabique de fréquence élevée en première position et un mot disyllabique de fréquence moyenne en seconde position.

Les cibles: les 21 mots trisyllabiques utilisés comme amorces, les 21 mots monosyllabiques correspondant aux composants enchâssés en première position, ainsi que les 21 mots disyllabiques correspondant aux composants enchâssés en seconde position. 105 items de remplissage ont été ajoutés. 3 listes ont été constituées, chacune contenant les 21 amorces reliées à une cible monosyllabique, disyllabique ou bien trisyllabique.

2.1.2.. Condition contrôle

Sujets: 10 étudiants de langue maternelle française.

Matériel

Les amorces: 63 mots trisyllabiques de fréquence basse, sans lien phonétique ou sémantique avec les cibles.

Les cibles: les mêmes que celles utilisées

dans la condition expérimentale. 63 items de remplissage ont été ajoutés.

Procédure et plan expérimental

Dans la condition expérimentale comme dans la condition contrôle, aucun sujet n'a vu la même cible ou entendu la même amorce deux fois. Il y a deux facteurs principaux: la Condition (2 niveaux: expérimentale et contrôle) et le Format de la cible (3 niveaux: monosyllabe, disyllabe et trisyllabe). La fin du signal sonore déclenche la mesure des temps de réaction (TR).

2.2. Résultats et discussion

Deux analyses de variance, par sujets et par items, ont été conduites. Dans la condition expérimentale comme dans la condition contrôle, les TR moyens étaient plus rapides pour les cibles monosyllabiques (484 ms, sd = 98 ms; 503 ms, sd = 120 ms) que pour les disyllabes (591 ms, sd = 130 ms; 565 ms, sd = 135 ms) ainsi que pour les trisyllabes (549 ms, sd = 137 ms; 618 ms, sd = 125 ms). Le facteur Format de la cible a un effet significatif, et cela par sujets et par items. Les TR sont fonction directe de ce format quand amorces et cibles ne sont pas reliées. Le traitement des disyllabes est toujours plus long quand elles sont reliées. L'effet du facteur Condition n'est pas significatif. Le calcul des effets d'amorçage (cf légende Fig. 1) montre que seul le traitement de la cible trisyllabique est facilité (+69 ms). Confrontés à des mots plurisyllabiques contenant des mots enchâssés, les auditeurs ont analysé le signal sonore comme étant un seul mot et non comme deux mots. Gow et Gordon (1995) ont appelé ce phénomène « avantage du mot le plus long ». Ces résultats confirment partiellement les prédictions faites à partir du modèle TRACE qui prévoit une activation du mot le plus long (le mot porteur) et une inhibition des mots enchâssés plus courts. Ils convergent également avec ceux obtenus par Frauenfelder et Peeters (1990) lors de simulations notamment quant à la tendance facilitatrice que nous obtenons pour le traitement du monosyllabe en position initiale. En effet, ces auteurs avaient mis en évidence une activation transitoire du monosyllabe enchâssé en première position. La faiblesse de l'effet facilitateur du monosyllabe pourrait être due à ce que son activation n'a pas eu un temps suffisant pour se propager avant qu'il soit recombiné avec l'information suivante. Si c'est le cas, nous faisons l'hypothèse qu'un intervalle de silence inséré entre les deux composants

enchâssés permettra la reconnaissance du monosyllabe.

3. EXPERIENCE 2

L'objectif de cette expérience est de « capturer » l'activation transitoire du monosyllabe enchâssé en position initiale en laissant suffisamment de temps après l'émission du monosyllabe pour que ce dernier soit activé et reconnu. Préalablement à cette expérience, nous avons conduit une expérience de discrimination de type AX dont l'objectif était d'évaluer l'intervalle de silence minimal qui amènerait les sujets à segmenter dans 50 % des cas les mots plurisyllabiques en leurs deux composants enchâssés. Une fois déterminé pour chaque item test, cet intervalle de silence (variant de 15 ms à 35 ms selon les items) a été introduit entre la fin acoustique du monosyllabe et le début acoustique du disyllabe en veillant à couper le signal lors de son passage par zéro. Cette manipulation du signal, en rendant inutilisables les indices de coarticulation entre les syllabes, devrait réussir à briser l'avantage du mot le plus long et ainsi permettre la reconnaissance des deux composants enchâssés, et particulièrement celle du monosyllabe qui bénéficie d'un temps supplémentaire pour que son activation se propage.

3.1. Méthode

Pour les conditions expérimentale et contrôle, la méthode est identique à celle de l'expérience 1 à la seule différence qu'un intervalle de silence a été introduit entre les deux composants de chacune des 21 amorces de la condition expérimentale.

Procédure et plan expérimental

La procédure et le plan expérimental sont identiques à ceux utilisés dans l'expérience 1.

3.2. Résultats et discussion

Dans la condition expérimentale, en moyenne, les décisions sur les monosyllabes (526 ms, sd = 110 ms) et les trisyllabes (511 ms, sd = 103 ms) tendent à être plus rapides que les décisions sur les disyllabes (550 ms, sd = 129 ms). Dans la condition contrôle, les TR étaient plus rapides pour une cible monosyllabique (448 ms, sd = 120 ms) que pour une cible disyllabique (488 ms, sd = 135 ms) ainsi que pour une cible trisyllabique (536 ms, sd = 125 ms). L'effet du facteur Format de la cible est significatif, et cela par sujets et par items. Seules les cibles monosyllabiques et disyllabiques présentent un effet d'amorçage inhibiteur significatif. (-78 ms et - 62 ms) (Fig. 2).

La présence de l'intervalle de silence a brisé l'avantage du mot le plus long puisque la facilitation du traitement des cibles trisyllabiques (+25 ms) n'est plus significative. Contrairement à nos hypothèses, aucun effet d'amorçage facilitateur n'est observé pour les deux composants enchâssés qui, au contraire, sont fortement inhibés. Il semble que le silence a permis à l'activation des deux mots enchâssés de se propager et que les effets inhibiteurs massifs observés sont le résultat des compétitions survenues entre les deux composants enchâssés activés et le mot trisyllabique enchâssant.

4. DISCUSSION GENERALE

Cette recherche avait pour objectif d'étudier la segmentation lors du traitement de mots ambigus et d'évaluer le jeu des compétitions entre hypothèses lexicales. L'expérience 1 montre un avantage du mot le plus long lorsque les auditeurs ont à traiter auditivement un mot plurisyllabique contenant des mots enchâssés. L'expérience 2 montre qu'une recherche lexicale en parallèle est certainement effectuée lors du traitement de séquences sonores ambiguës. Le candidat enchâssé en première position dans le mot porteur est probablement activé mais de façon transitoire.

Cependant, la présence d'indices pré-lexicaux (silence ou absence d'indices de coarticulation) ne suffit pas à déclencher une routine de segmentation. Les inhibitions massives observées dans l'expérience 2 suggèrent qu'une compétition importante a eu lieu entre d'une part, les deux composants enchâssés dont l'activation a eu le temps de se propager grâce à la présence de l'intervalle de silence, et d'autre part le mot enchassant qui, malgré tout, garde l'avantage d'être le mot le plus long, tendant de ce fait à recombiner la totalité de l'information acoustique.

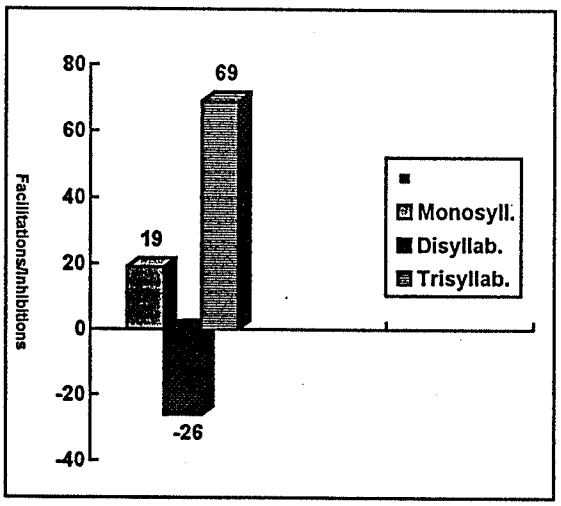


Figure 1: Différences en ms entre les temps de réponse dans la condition expérimentale et dans la condition contrôle pour chaque format de cibles (Exp. 1)

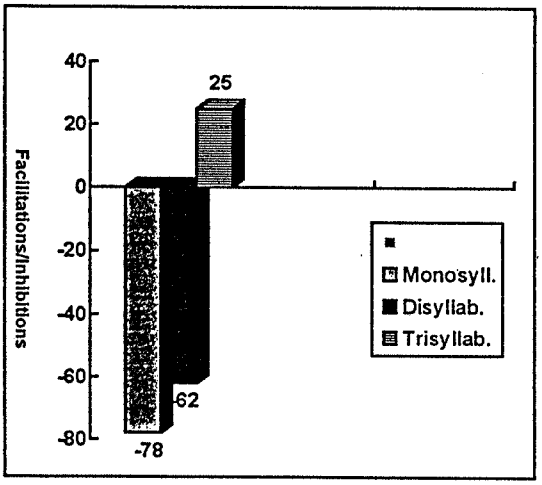


Figure 2: Différences en ms entre les temps de réponse dans la condition expérimentale et dans la condition contrôle pour chaque format de cibles (Exp. 2)

4. BIBLIOGRAPHIE

BANEL, M-H., BACRI, N. (1994), « On metrical patterns and lexical parsing in French », *Speech Communication*, vol. 15, pp. 115-126.

FRAUENFELDER, U., PEETERS, G. (1990), « Lexical segmentation in TRACE: An exercice in simulation », in G. Altmann (Ed.), *Cognitive models of speech processing*, Cambridge, MA: The MIT Press, pp. 50-86.

GOW, D. W., GORDON, P. C. (1995), « Lexical and prelexical influences on word segmentation: Evidence from priming », Journal of Experimental Psychology: Human Perception and Performance, vol. 2, pp. 344-359.

KLATT, D. H. (1980), « Speech perception: A model of acoustic-phonetic analysis and lexical access », In R.A Cole (Ed.), *Perception and production of fluent speech*, Hillsdale, N. J.: Erlbaum.

MARSLEN-WILSON, W. (1990), « Activation, competition, and frequency in lexical access », in G. Altmann (Ed.), *Cognitive models of speech processing*, Cambridge, MA: The MIT Press, pp. 148-172.

McCLELLAND, J., ELMANN, J. (1986), « The TRACE model of speech perception », *Cognitive Psychology*, vol. 18, pp. 1-86.

SHILLCOCK, R. (1990), « Lexical hypotheses in continuous speech », in G. Altmann (Ed.), *Cognitive models of speech processing*, Cambridge, MA: The MIT Press, pp. 24-49.

EFFET DE REPETITION INTERMODAL AVEC AMORÇAGE MASQUE

Elsa SPINELLI, Juan SEGUI et Jonathan GRAINGER

Laboratoire de Psychologie Experimentale, Universite René Descartes, CNRS URA 316
28 rue Serpente, 75006 Paris
Tel.: 40 51 98 65 - Fax: 40 51 70 85 - e-mail: spinelli@idf.ext.jussieu.fr

ABSTRACT

Two cross modal experiments have been carried out in order to test the hypothesis that phonological representations activated by visual masked primes are available to facilitate a subsequent auditory processing. In a lexical decision task, auditory targets are responded faster when preceded by briefly presented forward-masked identical primes than when preceded by non-related visual primes. With prime exposure duration of 29 ms, a repetition effect only occurs for high frequency words and rapid subjects whereas at 37 ms of prime exposure, the results yielded a repetition effect for all subjects although still restricted to high frequency words. This finding suggests that phonological representations activated by a briefly presented forward-masked prime can be available to facilitate the following processing of an auditorily presented target word, and underlines the necessity to take into account activation flow between orthographic and phonological representations in models of word recognition.

1. INTRODUCTION

La reconnaissance ou la production de mots exige qu'ils soient représentés et stockés en mémoire. Les mots étant spécifiés entre autres, par un code orthographique et un code phonologique, il paraît pertinent de savoir comment ces codes sont intégrés au sein du lexique et quelle est la nature des représentations utilisées pour l'accès dans différentes modalités sensorielles. En ce qui concerne la reconnaissance visuelle, l'idée que l'identification d'un mot implique non seulement l'activation de ses lettres mais également des sons qui lui correspondent a été démontrée à maintes reprises dans différentes tâches (Grainger & Ferrand, 1994; Perfetti, Bell & Delaney, 1988; Van Orden 1987; Jared & Seindenberg 1991); remettant ainsi en question la théorie de l'accès direct, point de vue selon lequel seules les représentations orthographiques sont à l'origine du contact avec les entrées lexicales (Baron, 1973). A la base de cette idée se trouvent les effets dits de *Pseudohomophonie* : en tâche de décision lexicale, les pseudohomophones comme *brane* (non-mot dont la prononciation est identique au mot *brain* en Anglais) requièrent plus de temps pour être jugés comme non-mots que des non-mots non pseudohomophones comme *slint* Coltheart, Davelaar, Jonasson & Besner (1977).

Plus récemment, Perfetti, Bell & Delaney (1988) ont montré que l'activation des représentations phonologiques lors du traitement des mots écrits s'effectuait de manière automatique. Ainsi, dans une expérience utilisant un paradigme de masquage rétroactif, ils observent que l'identification du mot anglais *made* est facilitée à la fois par les amorces *mard* et *mayd*, toutes deux orthographiquement similaires au mot cible mais avec une plus grande facilitation de *mayd*, pseudohomophone de made et donc seule capable d'activer les représentations phonologiques de la cible.

D'une manière analogue, il a été démontré que les représentations orthographiques peuvent être exploitées dans la reconnaissance auditive de mots (Fraunfelder, Segui & Dijkstra, 1990). Ainsi, Seindenberg et Tanenhaus (1979 Exp 2), ont montré dans une tâche auditive de détection de rimes, que la similarité orthographique entre les deux mots à comparer facilite la détection de rimes présentées auditivement. Les temps de latence sont plus courts lorsqu'il s'agit de détecter des rimes orthographiquement similaires (telles que *pie* et *tie* en anglais) par rapport aux rimes dont l'orthographe diffère (ex : *rye* et *tie* en anglais). De plus, la comparaison de mots qui ne riment pas mais dont l'orthographe est identique donne lieu à un allongement des temps de réponse (comme dans le cas de *leaf* et *deaf* en anglais).

Ces travaux suggèrent que lors du traitement d'un mot dans une modalité sensorielle, les sujets utilisent des représentations correspondant à l'autre modalité. Dans le souci de préciser la nature des représentations utilisées pour l'accès au lexique dans différentes modalités sensorielles, la présente recherche examine le rôle des

informations phonologiques activées par la présentation visuelle de mots, dans la reconnaissance de mots parlés. En effet, il s'agit de savoir si les représentations phonologiques activées par les mots présentés visuellement sont susceptibles d'être récupérées et utilisées lors du traitement auditif de ces mêmes mots.

Si c'est le cas, la présentation visuelle d'un mot devrait être à même d'influencer le traitement auditif de ce même mot présenté immédiatement après (effet de répétition).

2. EXPERIENCE 1

2.1 Méthode

2.1.1 Sujets

40 sujets de langue maternelle française recrutés parmi les étudiants de licence à l'université Paris V.

2.1.2 Matériel

Le matériel se compose de deux listes expérimentales de 128 items :

-32 mots monosyllabiques, non-homophones, dont :-16 mots de Haute Fréquence et 16 mots de Basse Fréquence qui constituent les cibles expérimentales ainsi que les amorces pour la condition de Répétition.

-32 mots contrôles appariés en longueur et en fréquence d'usage aux 32 mots expérimentaux, qui constituent les amorces pour la condition Non-Reliée.

-64 non-mots monosyllabiques qui respectent les contraintes phonotactiques de la langue française.

2.1.3 Procédure

La passation s'effectue individuellement devant un écran d'ordinateur. Le sujet effectue une tâche de décision lexicale sur les cibles auditives présentées dans un ordre aléatoire. Les amorces sont présentées visuellement selon la procédure de masquage suivante :

1-Affichage d'un pré-masque pendant 500 ms.
2-Affichage de l'amorce pendant 29 ms.
3-Affichage d'un post-masque et présentation de la cible auditive dans un casque. Le post-masque reste affiché à l'écran jusqu'à la réponse du sujet. Les items sont présentés dans deux conditions expérimentales : une condition Répétée (ex : joie- JOIE) et une condition Non Reliée (ex : vent-JOIE) contrebalancées afin que chaque sujet soit confronté à toutes les conditions expérimentales mais ne voie pas deux fois le même item cible.

Afin de maintenir l'attention du sujet sur le centre de l'écran (étant donné que les amorces visuelles ne sont pas perceptibles consciemment par le sujet), une tâche supplémentaire est demandée au sujet 12 fois au cours de l'expérience : identifier visuellement un mot présenté 50 ms et, comme les amorces visuelles, entre deux masques. Par ailleurs, à la fin de chaque expérience, on s'assure que les amorces visuelles n'ont pas été identifiées par le sujet en lui demandant d'effectuer une tâche d'identification des amorces visuelles présentées aléatoirement et dans les mêmes conditions que dans l'expérience.

2.2 Résultats

Table 1 : temps moyens de réponses correctes pour les mots en millisecondes en fonction des facteurs Fréquence et Relation.

	Répétée	Non reliée	Effet
H F	930	945	+15
B F	1064	1058	-6

Les analyses révèlent que le facteur Relation n'introduit pas d'effet significatif au niveau global. En revanche, on observe, pour les mots de Haute Fréquence une tendance à une diminution du temps de réponse pour les cibles présentées dans la condition répétée (930 ms) par rapport à la condition non reliée (945 ms). Cette différence de 15 ms observée pour les items de Haute Fréquence, bien que n'atteignant pas la significativité au seuil habituel de .05, mérite néanmoins notre attention. En effet, étant donné le caractère fugace des effets d'amorçage masqué, on peut supposer que seules les réponses les plus rapides (comme celles occasionnées par les mots de Haute Fréquence et pour les sujets rapides par exemple) permettent l'émergence d'un effet de répétition. Afin de confirmer ce point, des analyses plus fines ont été réalisées. Les 40 sujets ont été divisés en deux groupes de 20 sujets rapides et 20 sujets lents (selectionnés en fonction de leur moyenne globale).

Table 2 : Temps moyens de réponse des sujets en millisecondes en fonction des facteurs Fréquence, Relation et Rapidité des sujets.

	Sujets Rapides		Sujets Lents	
	Répétée	Non reliée	Répétée	Non reliée
H F	877	917	982	973
B F	1017	1011	1111	1103
moy	947	964	1047	1038

Pour le groupe des sujets rapides, on observe maintenant un effet de répétition qui se caractérise par des temps de réponse significativement plus courts, lorsque les amorces visuelles et les cibles auditives sont

identiques (947 ms) que lorsque amorces et cibles ne sont pas liées (964 ms), F(1-19)=6.32, p<.025. L'interaction entre les facteurs relation et fréquence étant significative (F(1-19)=15.84, p<.005), il semble clair que l'effet de répétition observé pour les sujets rapides est limité aux mots de Haute Fréquence, F(1-19)=14.4, p<.005. Enfin, le test d'identification visuelle des amorces effectué par les sujets révèle un taux d'identification quasiment nul..

Cet effet de répétition intermodal obtenu chez les sujets les plus rapides et pour les mots de Haute Fréquence semble indiquer que l'activation des informations phonologiques par la présentation visuelle des mots facilite le traitement auditif ultérieur de ces mêmes mots. Cependant, l'effet d'amorçage demeure fragile étant donné que seules les réponses les plus rapides (produites par les mots de Haute Fréquence et les sujets rapides) sont affectées par l'amorce visuelle. Ce genre d'effet est très ténu, et il pourrait disparaître si le traitement de la cible demande trop de temps (comme dans le cas des mots de Basse Fréquence ou des mots trop longs par exemple). Afin de procurer au sujet davantage d'information activée par l'amorce visuelle, la durée de présentation des amorces a été augmentée à 37 ms dans l'expérience 2.

3. EXPERIENCE 2

3.1.Méthode

3.1.1. Sujets

Les sujets sont 26 étudiants de langue maternelle française recrutés parmi les étudiants de licence à l'université Paris V et n'ayant pas participé à l'expérience précedente.

3.1.2 Matériel

Le matériel est le même que celui de l'expérience 1.

3.1.3 Procédure

Les procédures sont les mêmes que dans l'expérience 1 à ceci près que la durée de présentation de l'amorce est de 37 ms. A la fin de la passation de l'expérience, le sujet doit identifier les amorces présentées visuellement pendant 37 ms.

3.2.Résultats

Table 3 : Temps moyens de réponses correctes pour les mots en millisecondes, en fonction des facteurs Fréquence et Relation.

	Répétée	Non reliée	Effet
H F	927	954	+27
B F	1053	1046	-7

Comme pour l'expérience 1, les analyses révèlent que le facteur Relation n'introduit pas d'effet significatif au niveau global.

Pour ce qui est du facteur Relation, on observe que pour les mots de Haute Fréquence, les temps de réponse sont significativement plus courts lorsque les amorces visuelles et les cibles auditives sont identiques (927 ms) que lorsque amorces et cibles ne sont pas liées (954 ms), F(1-24)=11.8, p<.005.

Les résultats des deux expériences paraissent cohérents entre eux et semblent indiquer qu'avec un temps de présentation plus important de l'amorce visuelle (37 ms au lieu de 29 ms), on observe pour les mots de haute fréquence, un effet de répétition généralisable à l'ensemble des sujets. Il nous est possible d'interpréter ces résultats en terme d'activation et d'écarter l'hypothèse d'une stratégie des sujets basée sur l'identification des amorces étant donné que le post-test d'identification des amorces révèle, comme dans l'expérience 1 un taux d'identification quasiment nul.

4. DISCUSSION GENERALE

Dans la présente recherche, notre but était de savoir si les représentations phonologiques activées par la présentation visuelle et masquée d'un mot peuvent intervenir lors d'un traitement auditif ultérieur. Dans notre première expérience, on constate qu'à 29 ms de présentation de l'amorce, l'effet facilitateur des amorces présentées en condition de répétition n'apparaît que pour les mots de haute fréquence et les sujets les plus rapides.

En revanche, dans notre deuxième expérience où les amorces sont présentées pendant 37 ms, les amorces répétées facilitent le traitement des cibles auditives et ce, pour l'ensemble des 26 sujets, bien que l'effet demeure restreint aux mots de haute fréquence.

Les résultats de ces deux expériences sont cohérents entre eux et suggèrent qu'un transfert d'activation des unités orthographiques vers les unités phonologiques correspondantes est possible malgré la fragilité de tels effets d'amorçage. La fragilité de ces résultats soulève deux remarques : La première, méthodologique, sur la difficulté à mettre en évidence des effets d'amorçage masqué avec un paradigme de décision lexicale auditive. Effectivement, le traitement des cibles étant un traitement auditif, il requiert plus de temps qu'un traitement visuel puisqu'il se déroule de manière séquentielle. Etant donné le caractère éphémère des effets d'amorçage rapide et masqué, il y a un risque de

déperdition de l'effet des amorces si le traitement des cibles requiert trop de temps. Nous avons pallié à cette difficulté en ne choisissant que des mots monosyllabiques (donc de courte durée); cependant il serait intéressant d'utiliser des mots plus longs en présentant l'amorce visuelle lorsque le traitement de la cible est en cours.

La deuxième remarque concerne l'absence de facilitation des amorces répétées sur le traitement des cibles de basse fréquence. En effet, l'effet de répétition n'apparaît pas pour les mots de basse fréquence, et ceci quelle que soit la durée de présentation de l'amorce. Si l'on considère que le seuil de réponse pour les mots de basse fréquence est plus élevé que le seuil de réponse pour les mots fréquents comme dans les modèles d'activation comme celui des logogens de Morton (1979) ou si l'on estime que les détecteurs de mots rares ont un niveau d'activation au repos (ou de base) plus bas que les détecteurs de mots fréquents comme le postule le modèle d'activation interactive de Mac Clelland et Rumelhart (1981), on admet communément que la quantité d'information sensorielle nécessaire au déclenchement de l'identification d'un mot est plus grande pour les mots rares que pour les mots fréquents. Ainsi, l'absence d'effet pour les mots de Basse Fréquence peut être interprétée comme étant due à une activation insuffisante de leurs représentations dans les conditions de répétition.

Ainsi, la présente recherche montre que les représentations phonologiques activées par les amorces visuelles de haute fréquence peuvent également être utilisées lors du traitement auditif subséquent des cibles. Ceci ne remet pas en cause les théories qui postulent l'existence de deux lexiques distincts, l'un contenant les représentations orthographiques et l'autre les représentations phonologiques, comme celle de Morton (1979) qui envisage des logogènes spécifiques aux différentes modalités.

Nos résultats permettent seulement de suggérer que les mécanismes responsables de la reconnaissance du langage écrit et parlé ne sont pas « imperméables » et qu'il existe des interrelations entre eux. Afin de rendre compte de l'effet facilitateur de la présentation visuelle d'un mot sur le traitement auditif subséquent de ce même mot mis en évidence dans nos expériences, il serait pertinent d'envisager un lien entre les lexiques orthographique et phonologique par lequel le flux d'activation pourrait circuler. D'un point de vue méthodologique, la nouvelle application du paradigme d'amorçage masqué à l'étude de la reconnaissance des mots parlés peut fournir un nouvel outil pour les études futures en inter-modalité.

5. BIBLIOGRAPHIE

Baron, J. (1973). Phonemic stage not necessary for reading. *Quaterly journal of Experimental Psychology*, **25**, 241-246.

Coltheart, M., Davelaar, E., Jonasson, J. T. & Besner, D. (1977). Acces to the internal lexicon, *in* S. Dornic (ed.), *Attention and performance VI*, New York, Academic Press.

Fraunfelder, U. H., Segui, J. & Dijkstra, T. (1990). Lexical effects in phonemic processing : Facilitatory or inhibitory ?, *Journal of experimental psychology: Human Perception and Performance*, **16**, **1**, 77-91.

Grainger, J. & Ferrand, L. (1994). Phonology and orthography in visual word recognition : Effects of masked homophones primes. *Journal of Memory and Language*, **33**, 218-233.

Jared, D. & Seindenberg, M. S. (1991). Does word identification proceed from spelling to sound to meaning ? *Journal of experimental psychology : General* **120**, 358-394.

Mac Clelland, J. L. & Rumelhart, D E. (1981). An interactive activation model of context effects in letter perception : Part I. An account of basic findings, *Psychological review*, **88**, 375-407.

Morton, J. (1979). Facilitation in word recognition : Experiments causing change in the logogen models. In P. A. Kolers, M. E. Wrolstad & H. Bouma (eds.), *Processing of visible language*, **1**, 259-268. New York : Plenum.

Perfetti, C. A., Bell, L. C. & Delaney, S. M. (1988). Automatic (prelexical) phonetic activation in silent word reading : Evidence from backward masking, *Journal of Memory and Language*, **27**, 59-70.

Seidenberg, M. S. & Tanenhaus, M. K. (1979). Orthographic effects on rhyme monitoring, *Journal of Experimental Psychology : Human Learning and Memory*, **5(6)**, 546-554.

Van Orden, G. C. (1987). A ROWS is a ROSE : Spelling, sound and reading, *Memory and Cognition*, **15**, 181-198.

ETUDE PERCEPTIVE DU DÉPHASAGE ENTRE LES GESTES LABIAL ET LINGUAL EN SYNTHESE ARTICULATOIRE

Valérie Pasdeloup et Paul Jospa

Institut des Langues Vivantes et de Phonétique (CP 110), ULB,
av. Franklin Roosevelt 50, 1050 Bruxelles, Belgique. E-mail : valpasde@ulb.ac.be

ABSTRACT

This paper describes a perceptual experiment which was run to study the difference in phase between lip and tongue movements. Realistic, irrealistic and no difference in phase have been studied in V_1V_2 sequences consisting of protruded and non-protruded vowels. This was done within a gestural model apply to articulatory synthesis. The results suggest that subjects show preferences for realistic difference in phase between articulators rather than irrealistic one.

1. INTRODUCTION

Le phasage du mouvement des articulateurs ou de régions spécifiques du conduit vocal est pris en compte dans de nombreuses études portant notamment sur la production (Nittrouer & al., 1988) et la phonologie articulatoire (Browman & al., 1986). Nous nous proposons ici d'étudier l'effet perceptif du déphasage du mouvement labial par rapport au mouvement des régions contrôlées par la langue. en synthèse de segments V_1V_2 comportant une voyelle protruse et une non protruse. Cette étude est réalisée dans le cadre d'un modèle de génération de transitions articulatoires (Jospa & al., 1995) supportant les notions de cibles articulatoires, de coactivation (Nittrouer & al., 1988), de partitions gestuelles (Browman & al., 1986) et de superposition du geste consonantique à un continuum vocalique (Carré & al., 1992). Les profils articulatoires (formes du conduit) sont exprimés selon le modèle à régions distinctives (DRM) (Carré & al., 1992). Les paramètres "articulatoires" (distances sagittales des 8 régions et longueur du conduit) sont mesurées relativement à la position neutre du conduit. Une cible articulatoire Y_i^* est définie par l'ensemble des 9 paramètres articulatoires associés à un phonème i. Le modèle est gouverné par des fonctions d'activation des cibles, de nature sigmoïdale, qui sont contrôlées par les paramètres suivants (voir fig.1): t_0=temps central de transition, Δ = déphasage du geste labial vis à vis du geste lingual, G = temps de séparation entre deux fonctions d'activation successives (G = 0 dans notre étude, et d = durée effective d'établissement d'une cible. La loi de composition des cibles donnant lieu au mouvement de transition est:

$$y_k^{V_1V_2}(t) = \alpha_{V_1,k}(t).h_{V_1}.Y^*_{V_1,k} + \alpha_{V_2,k}(t).h_{V_2}.Y^*_{V_2,k}$$

avec: $0 \le \alpha_{V_1,k}(t), \alpha_{V_2,k}(t) \le 1$,

où: l'indice k désigne le paramètre articulatoire (région linguale ou labiale), les $Y^*_{V_1,k}$ et $Y^*_{V_2,k}$ sont deux cibles invariantes pour la région k, les h_{V_1}, h_{V_2} (=1 dans notre cas) sont des facteurs d'amplification des cibles et les $\alpha(t)$ sont les fonctions d'activation des cibles de forme sigmoïdale.

Lors d'une évaluation acoustique et perceptive préliminaire de ce modèle gestuel, différents types de "patrons de coactivation" V_1V_2 ont été testés. Des séquences V_1V_2 de bonne qualité ont pu ainsi être obtenues avec des patrons d'activation V_1V_2 de forme simple : pour les parties stables des voyelles avec un degré d'activation maximal et pour la transition avec des pentes sigmoïdales et symétriques, c'est-à-dire qui se chevauchent en leur milieu. L'activation d'une cible neutre intermédiaire durant la transition V_1V_2, lorsque les pentes se chevauchent peu, ne semble pas améliorer la qualité. La coactivation de la voyelle suivante ou précédente durant la partie stable d'une voyelle n'apparaît pas nécessaire ; celle-ci donne lieu à une trop grande modification des valeurs de formants et n'améliore pas la qualité de la transition.

Dans une première version du modèle gestuel, les patrons d'activations étaient générés de telle sorte qu'une modification dans l'amplitude et dans le timing de l'activation affectait de la même manière toutes les régions du conduit. Ceci ne nous permettait pas de rendre compte des effets d'anticipation et de persévération de certains mouvements articulatoires relativement à d'autres.

Afin de tester les effets d'anticipation et de persévération, nous avons introduit dans le modèle gestuel des déphasages temporels entre différents groupes de régions. De façon à rester au plus proche des articulateurs, nous

avons groupé d'une part la région des lèvres (R8) et la longueur du conduit, afin de rendre compte du mouvement de protrusion labiale, et d'autre part les autres régions (R2-7) afin de rendre compte du mouvement lingual. Nous avons émis l'hypothèse que des déphasages temporels entre ces deux groupes de régions, c'est-à-dire lèvres-langue, devraient améliorer la qualité de séquences VVp constituées d'une voyelle protruse (Vp) et d'une voyelle non protruse.

Dans le cas d'une séquence VVp, comme é-o, on s'attend à une anticipation de la protrusion sur le mouvement de la langue, donc à un déphasage négatif de l'activation de la région des lèvres par rapport à l'activation des autres régions; dans le cas d'une voyelle protruse suivie d'une voyelle non protruse comme o-é, on s'attend à une perséveration de la protrusion sur le mouvement de la langue, donc à un déphasage positif.

2. EXPÉRIENCE

L'objectif de la présente expérience est l'étude perceptive du déphasage du mouvement labial par rapport au mouvement lingual, dans le contexte VVp et VpV. Trois types de déphasages ont été étudiés : un déphasage positif de l'activation de la région des lèvres et de la longueur du conduit par rapport à l'activation des autres régions, un déphasage négatif et aucun déphasage.

Méthode

Matériels. Sept voyelles, 3 non-protruses (a, è, i) et 4 protruses (eu, o, ou, u), produites par un locuteur francophone, ont été sélectionnées. Les formes du conduit correspondantes ont été stockées après avoir été exprimées en terme du Modèle à Régions Distinctives. Vingt-quatre séquences VVp et VpV ont été ainsi constituées, 12 dans un ordre et 12 dans l'ordre inverse.

La durée des séquences V_1V_2 était de 450ms et celle des transitions de 100ms. Trois types de déphasage des lèvres par rapport à la langue ont été synthétisés : un déphasage positif de +50ms, un déphasage négatif de -50ms et un déphasage nul. L'amplitude de l'activation était fixée à son maximum durant les parties stables des voyelles et il n'y avait pas de coactivation d'une autre voyelle durant les parties stables des voyelles. Les patrons d'activation étaient identiques pour les deux groupes de régions lèvres+conduit/langue.

Les séquences V_1V_2 ont été présentées par paires, avec une durée d'une seconde entre chaque membre de la paire. Les sujets devaient choisir dans chaque paire la séquence qu'ils préféraient. Pour chacune des 24 séquences V_1V_2, 3 paires et leurs symétriques, soit 6 paires, ont été créées :

- déphasage négatif et déphasage positif,
- déphasage négatif et déphasage nul,
- déphasage positif et déphasage nul.

Cent quarante-quatre paires de séquences V_1V_2 ont été ainsi constituées : 3 V X 4 Vp = 12 séquences, 2 positions initiale ou finale pour Vp, 3 types de paires, 2 ordres de présentation.

Sujets. 20 sujets francophones ont passés le test.

Procédure. Les sujets ont été testés deux par deux et entendaient les paires de séquences V_1V_2 avec un casque. La liste des séquences était écrite dans un carnet. La tâche du sujet consistait pour chaque paire de séquences à choisir celle qu'il préférait et à l'indiquer par une croix. Les sujets entendaient une nouvelle paire de séquences après 3 secondes. Chaque paire était présentée 2 fois. Les paires étaient présentées dans un ordre pseudo-aléatoire et 288 paires étaient présentées aux sujets. La session était précédée d'une phase d'entraînement de 10 paires. Le test durait environ 25mn.

3. RÉSULTATS

Nous avons réalisé une analyse statistique de régression logistique qui nous a permis de prendre en compte le fait que les réponses étaient binaires. Dans cette analyse, nous avons considéré les 12 items (séquences V_1V_2), le facteur position (Vp initiale et finale), le facteur déphasage (3 types de paire). Le facteur ordre de présentation des séquences dans la paire a été considéré comme un facteur aléatoire.

Les résultats sont les suivants : Le facteur items est significatif à .00001. Les préférences des sujets sont par conséquent différentes selon les séquences présentées. Le facteur déphasage est significatif à .00001. Les préférences des sujets sont différentes quand on leur présente les 3 types de paires (déph. +/-, +/0, -/0). Le facteur ordre de présentation (Vp initiale ou finale) est significatif à .0001. Ce qui signifie que les

préférences des sujets sont différentes lorsqu'on leur présente une paire de séquence constituée des mêmes voyelles, mais présentées dans l'ordre inverse. Les interactions entre les facteurs déphasage et items, et entre items et position sont toutes deux significatives à .00001.

De plus, l'analyse logistique nous a permis de tester si les réponses des sujets étaient ou non aléatoires. Si nous considérons l'ensemble des réponses, elles n'ont pas une distribution aléatoire. Ce résultat est significatif à .00001. Si nous considérons les réponses pour chaque type de séquence V_1V_2, les préférences n'étaient pas aléatoires pour toutes les séquences a+Vp, Vp+a, i+Vp, Vp+i et pour o-è, è-ou, è-u, u-è. Les réponses étaient aléatoires pour è-eu, eu-è, è-o, ou-è.

Afin de pouvoir comparer les résultats de séquences symétriques VVp et VpV comme é-o et o-é, nous avons utilisé d'autres termes pour désigner les déphasages positifs et négatifs. Ceux-ci ne prennent pas en compte le fait que le déphasage est positif ou négatif, mais *le délai (D) entre le mouvement labial et le milieu de la partie stable de Vp.* Prenons par exemple la séquence VVp é-o, l'anticipation du mouvement labial correspond à un déphasage négatif ; dans ce cas, D correspond au délai le plus long pour les 3 cas de déphasages ; nous l'appellerons *D long* . Ce délai est plus long que le délai correspondant à un déphasage nul que nous appellerons *D 0.* Et ce dernier est lui-même plus long que le délai nécessaire à un déphasage positif, qui sera appelé *D court.* Notons que ces termes nous permettent de nommer *D long*, à la fois le déphasage négatif dans le cas où Vp est en position V2 et le déphasage positif dans le cas où Vp est en position V1. Notre hypothèse étant que *D long* correspond à un déphasage "réaliste" et *D court* à un déphasage "irréaliste", nous nous attendons à ce que les sujets expriment une préférence pour *D long*, lorsqu'on leur a présenté les paires *D long/court* et *D long/0*, et pour *D 0* pour la paire *D 0/court.*

Tableau des préférences moyennes pour les 3 types de paires de séquences

	Vp1	Vp2	délai long/court préférence: long	délai 0/court préférence: 0	délai long/0 préférence ?	long/0 % préf.	trajet. acoustique la plus courte
		eu	oui	oui	0	60	0
	eu		oui	oui	0	55	0
		o	oui	oui	long	55	long and 0
avec a	o		oui	oui	0	52,5	long et 0
		ou	oui	oui	pas de préf.	50	long et 0
	ou		oui	oui	0	55	long et 0
		u	oui	oui	0	87,5	0
	u		oui	oui	0	72,5	0
		eu	aléatoire	aléatoire	0	58,75	0
	eu		aléatoire	aléatoire	0	55	0
		o	aléatoire	aléatoire	0	52,5	toutes ~ =
avec è	o		aléatoire	aléatoire	0	55	toutes ~ =
		ou	aléatoire	aléatoire	0	58,75	toutes ~ =
	ou		aléatoire	aléatoire	0	55	toutes ~ =
		u	aléatoire	oui	0	91,25	0
	u		oui	oui	0	91,25	0
		eu	oui	oui	pas de préf.	50	toutes ~ =
	eu		oui	oui	long	55	toutes ~ =
		o	oui	oui	0	52,5	long et 0
avec i	o		oui	oui	long	53,75	long et 0
		ou	oui	oui	long	58,75	long
	ou		oui	oui	long	61,25	long
		u	oui	oui	0	53,75	=
	u		oui	oui	0	52,5	=

Le tableau ci-dessus présente les préférences moyennes des sujets pour chaque séquence V_1V_2 pour chacune des 3 paires présentées, c'est-à-dire délais long/court, 0/court et long/0. En général, bien que le facteur présentation soit significatif, on observe les mêmes patrons de réponse pour

une séquence V_1V_2 et son symétrique V_2V_1. C'est le cas par exemple pour la séquence i-ou, où *D long* (ici un déphasage négatif) est préféré à *D 0*, et ou-i où *D long* (ici un déphasage positif) est préféré à *D 0*.

De plus, les sujets en général n'ont jamais préféré dans une paire le déphasage qui contredisait nos hypothèses. Pour la paire *D long/court*, les sujets ont préféré *D long* ou on répondu au hasard, et pour la paire *D 0/court*, les sujets ont préféré *D 0* ou ont répondu au hasard. Comme nous pouvons le voir dans le tableau, dans aucun cas les sujets n'ont montré une préférence pour des mouvements articulatoires "irréalistes", c'est-à-dire par exemple pour une anticipation du mouvement des lèvres sur la langue dans le contexte VpV. Les sujets ont montré des préférences pour des mouvements articulatoires "réalistes" ou bien n'ont montré aucune préférence, comme c'est le cas pour la plupart des séquences è+Vp et Vp+è. Par exemple pour les séquences i-ou, la préférence moyenne pour un déphasage positif lèvres/langue (persévération de la protrusion) n'est jamais supérieure à la préférence moyenne pour un déphasage négatif (anticipation de la protrusion).

En ce qui concerne les préférences entre *D long* et *D 0*, pour 71% des séquences les sujets préfèrent *D 0* à *D long* ; cependant, pour 4 séquences seulement, a-u, u-a, é-u et u-è cette préférence est marquée et dépasse 70%. Dans les autres cas, cette préférence est inférieur à 60%. Pour 8% des séquences, a-ou et i-eu, les sujets n'ont pas eu de préférence. Pour 21% des séquences, a-o, eu-i, o-i, i-ou et ou-i, les sujets préfèrent *D long* à *D 0* ; cependant, ces préférences ne sont pas très marquées.

DISCUSSION

Notre objectif était de tester perceptivement le déphasage du geste de protrusion labiale par rapport au geste lingual dans des séquences VVp et VpV. Il apparaît en premier lieu que les sujets n'ont jamais préféré un déphasage correspondant à un mouvement articulatoire "irréaliste" : anticipation labiale dans VpV et persévération labiale dans VVp. Par contre entre une absence de déphasage et un déphasage "réaliste", les sujets ont pour les 3/4 des séquences préféré l'absence de déphasage, mais cette préférence n'est pas très marquée. Les préférences dépendent du type de séquences V_1V_2 (le facteur items est significatif dans l'analyse logistique). Ainsi pour les séquences avec la voyelle i (i+Vp et Vp+i), dans la moitié des séquences les sujets ont préféré un déphasage "réaliste" à l'absence de déphasage.

Afin d'essayer de rendre compte de cette variabilité entre les items, nous avons étudié les caractéristiques acoustiques de ces séquences, et plus particulièrement les trajectoires formantiques dans le plan F1/F2. Si l'on observe les trajectoires durant les transitions, on remarque que pour toutes les séquences les sujets n'ont jamais préféré les trajectoires les plus longues. Les sujets ont montré leur préférence pour la trajectoire la plus courte, ou pour une des trajectoires les plus courtes quand elles étaient assez proches. Dans le cas des paires *D long/court*, où les sujets ont préféré *D long*, la trajectoire formantique correspondant à *D long* est généralement plus courte que celle correspondant à *D court*.

Remerciement : cette recherche a été financée par un contrat de la CE : programme Science (n°SC1*-CT92-0786).

RÉFÉRENCES

Nittrouer, S. & al. (1988) "Patterns of interarticulator phasing and their relation to linguistic structure", *JASA*, **84**, 1653-1661.

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

Jospa, P.,George, M. & Soquet, A. (1995). "A gestural production model built by acoustic-articulatory inversion of formant frequencies". Proc. ICPhS 95 Stockholm, **2**, 446-449.

Carré, R. & Mrayati, M. (1992): "Distinctive regions in acoustic tubes. Speech production modelling", J. Acoustique, 5, 141-159.

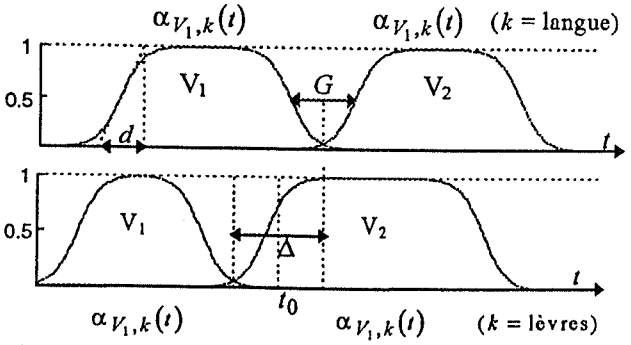


Figure 1: Définition des paramètres de contrôle des fonctions d'activation: $\alpha_{V_1,k}(t)$ et $\alpha_{V_2,k}(t)$.

IDENTIFICATION DES VOYELLES À PARTIR DU BRUIT DES OCCLUSIVES

Anne BONNEAU

CRIN–CNRS & INRIA Lorraine, Bâtiment LORIA, BP 239, 54506 Vandœuvre-lès-Nancy

Tél: 83 59 20 80 - Fax : 83 41 30 79 - email : bonneau@loria.fr

ABSTRACT

This paper deals with the perception of vowels from French stop bursts. The corpus was made up of 90 stimuli of 20-25 ms extracted from natural CVC and CV monosyllabic words. The syllables combined the three initial voiceless stops /p,t,k/ with the three vowels /i,a,u/. In order to cut off all traces of vocalic segment, bursts whose duration was too short have been lengthened. Eight native speakers of French served as listeners in the experiment. Results showed that a burst onset which did not contain any traces of vocalic segment provided substantial vocalic information (the overall identification rate was 80%). The vowel /i/ was clearly identified from /t/ and /k/, and the vowel /u/ very clearly identified from /k/. The vowel /a/, with high identification rate, was often chosen by listeners when they could not identify the color of the vowel. Some of the vocalic cues provided by the burst have been investigated.

1. INTRODUCTION

Le bruit d'explosion est à lui seul, c.a.d sans l'apport d'autres sources d'information, un indice fiable du lieu d'articulation des occlusives (cf Stevens et Blumstein, 1979; Kewley-Port *et al.*, 1983; Krull, 1990; entre autres). La connaissance de la voyelle subséquente, -que son identité soit révélée à l'auditeur, ou que la voyelle, dépourvue de transitions, soit ajoutée au signal de bruit-, n'accroît que faiblement l'intelligibilité de la consonne (Repp et Lin, 1989; Bonneau *et al.*, 1996). L'amélioration apportée par la connaissance de l'identité du contexte à l'identification des occlusives n'est peut-être pas pour autant négligeable. Le bruit lui-même fournit des informations sur l'identité de la voyelle subséquente. Si le rôle joué par cette source d'information est difficile à cerner, il nous semble néanmoins précieux. C'est pourquoi nous nous proposons d'étudier ici la nature et l'importance de l'information vocalique fournie par le bruit d'explosion.

Selon (Winitz *et al.*, 1972), les voyelles /i,a,u/ peuvent être identifiées à l'écoute du bruit d'explosion des consonnes occlusives sourdes aspirées de l'anglais, avec les taux de reconnaissance suivants : 90% pour /i/, 54% pour /a/ et 50% pour /u/ (l'identité de la consonne était connue des auditeurs pendant le test). Repp et Lin (1989) ont montré que la classe de la voyelle, définie comme la voyelle et ses deux plus proches voisines, pouvait être identifiée à partir de la partie transitoire du bruit d'explosion des occlusives extraites de syllabes CV chuchotées (taux d'identification : 61%). Les voyelles n'ont pas été identifiées de manière aussi précise, mais avec un taux bien supérieur à la chance (28% en moyenne pour un corpus comprenant 11 voyelles distinctes). Dans l'expérience de Cullinan et Tekieli (1979), les auditeurs ont pu identifier certains traits vocaliques dès les 10 ou 20 premières millisecondes des syllabes occlusive-voyelle. Avec trente millisecondes de bruit, les sujets ont identifié correctement le trait avant-arrière dans 90% des cas, et le trait d'ouverture dans 60% des cas. Dans les trois expériences présentées ci-dessus, les voyelles antérieures fermées /i/ et /I/ ont été les mieux identifiées. Nous avons procédé à une nouvelle expérience de perception afin de vérifier si le bruit des occlusives françaises, dépourvu de segment vocalique, est également porteur d'information sur la voyelle subséquente.

2. PROTOCOLE EXPÉRIMENTAL

Nous avons repris le corpus et les stimuli d'une expérience concernant l'identification des occlusives (Bonneau *et al.*, 1996). Ce corpus comprend des mots monosyllabiques constitués de suites *occlusive sourde-voyelle (/i/, /a/ ou /u/)* parfois suivies d'une consonne. Chacune des neuf combinaisons *occlusive-voyelle* apparaît deux fois dans le corpus (pas toujours dans le même mot) et chaque paire de syllabes similaires est prononcée par cinq locuteurs masculins (soit un total de 90 mots). La durée des stimuli, fixée à 25 ms environ lors de l'expérience précédente, a été conservée car elle répond à une double exigence : la durée doit être suffisamment brève pour ne contenir que le début de la syllabe et suffisamment longue pour que les stimuli soient bien perçus. En outre, nous désirons pouvoir effectuer des comparaisons entre les résultats des deux expériences. Les stimuli utilisés ne contiennent aucune trace de segment vocalique. En effet, nous avons allongé les bruits trop brefs en dupliquant la fin du bruit, c.a.d essentiellement le faible bruit de friction de quelques labiales (un peu plus d'un tiers des labiales ont été allongées, et seulement trois dentales).

Huit auditeurs français ont participé à l'expérience. Les sujets ont écouté les stimuli à l'aide d'un casque Sennheiser HD520 II dans une pièce calme. Le volume sonore a été réglé à un niveau d'écoute confortable. Afin de familiariser les auditeurs avec les stimuli, nous les avons soumis à une séance d'apprentissage. Le corpus d'apprentissage était constitué de 27 répétitions différentes des mots du corpus de test. Les 9 combinaisons *occlusive sourde-voyelle (/i/, /a/ ou /u/)* ont été prononcées par trois locuteurs masculins. À partir de chaque mot du corpus d'apprentissage, nous avons généré quatre stimuli de longueur différente : le bruit plus 2/3 de la voyelle, le bruit plus 1/3 de la voyelle, le bruit seul, les stimuli de 25 ms dépourvus de segment vocalique. Les auditeurs ont écouté les 108 (27x4) stimuli tout en lisant leur identité. Ils ont ensuite subi un test de perception avec les stimuli brefs (de 25 ms), présentés en ordre aléatoire, puis ont écouté à nouveau les stimuli mal identifiés tout en lisant la bonne réponse. L'apprentissage a duré moins de 12 mn.
Le corpus de test a été présenté trois fois aux auditeurs, dans un ordre aléatoire chaque fois différent. Nous avons demandé aux sujets de choisir leur réponse parmi les trois voyelles /i,a,u/, puis nous leur avons donné des feuilles de papier comportant trois colonnes, une pour chaque voyelle, en leur disant de mettre une croix dans la colonne correspondant à leur réponse. Une pause de 10 minutes a été intercalée entre l'apprentissage et le premier test d'une part, et les deux derniers tests d'autre part. Les stimuli ont été présentés par blocs de 9 stimuli, avec une pause de 10 secondes entre chaque bloc et une pause de 4 secondes entre chaque stimulus. Chacun des huit auditeurs a donné trois réponses pour chaque stimulus, ce qui représente 24 réponses par stimulus et 2160 (24x90) réponses au total.

3. RÉSULTATS

Le taux moyen de reconnaissance est de 80%. Nous avons soumis les données à une analyse de variance, ANOVA, afin d'examiner les effets de deux facteurs : la consonne (trois niveaux) et la voyelle (trois niveaux). L'effet de la voyelle n'est pas significatif ($p = 0.14$). Bien que les scores moyens d'identification de chaque voyelle soient bien distincts (84% pour /a/, 80% pour /i/, et 75% pour /u/), les variations autour de ces moyennes sont relativement importantes. En revanche, l'effet de la consonne est significatif ($p = 0.02$). Le test de Scheffé montre que les voyelles sont mieux identifiées à partir du bruit de la consonne /k/. Ce résultat était attendu puisque l'articulation de /k/ est très influencée par celle de la voyelle subséquente. Les taux d'identification par consonne sont de 88% pour /k/, 79% pour /p/ et 76% pour /p/. L'interaction entre la consonne et la voyelle ($p < 0.0001$) est très significatif.

Il semble que les sujets aient choisi la réponse /a/ en l'absence de timbre vocalique distinct. Ils ont en effet unanimement déclaré ne pas reconnaître le timbre de /a/, mais ont parfois clairement distingué les timbres de /i/

Table 1: Matrices de confusion. Scores (%) pour chaque syllabe et pour chaque voyelle. Les colonnes indiquent les réponses.

	i	a	u
pi	**63**	31	5
ti	**90**	9	1
ki	**86**	13	1

	i	a	u
pa	1	**85**	14
ta	3	**94**	3
ka	21	**74**	5

	i	a	u
pu	0	9	**91**
tu	2	61	**37**
ku	0	1	**99**

	i	a	u
i	**80**	18	2
a	9	**84**	6
u	1	24	**75**

et de /u/. Cela explique probablement le grand nombre de confusions de /i/ vers /a/ (18%), et de /u/ vers /a/ (24%). Les confusions inverses de /a/ vers /i/ (9%) et de /a/ vers /u/ (6%) sont nettement moins nombreuses, les confusions entre /i/ et /u/ sont rares. La réponse /a/ est donc attribuée à la fois au contexte /a/ et aux contextes /i/ et /u/ peu distincts.

Pour ce qui concerne l'identification des traits vocaliques, le faible nombre de confusions entre /i/ et /u/ démontre la très bonne identification du trait avant-arrière.

Si on ne tient pas compte du contexte /a/, qui ne semble pas identifié en tant que tel, nous remarquons que les voyelles sont mieux identifiées quand la consonne et la voyelle subséquente ont des articulations voisines (cf /u/ sur /k/, 99%, /i/ sur /t/, 90%, /i/ sur /k/, 86%).[1] Rappelons que /k/ est vélaire devant /u/, voyelle vélaire et fermée, qui obtient un taux d'identification quasi-optimal, et que /k/ est post-palatal devant la voyelle palatale /i/. Les expériences précédentes (Repp et Lin, 1989; Winitz *et al.*, 1972; Cullinan et Tekieli, 1979) mettent toutes en évidence la grande intelligibilité du contexte /i/. En revanche, /u/ est mal identifié à partir de /t/, les articulations des deux sons étant probablement trop éloignées pour que la voyelle soit coarticulée au relâchement de la consonne. La mauvaise identification de /u/ à partir de /t/ est du reste une constante dans toutes les expériences sur

[1]La proximité des articulations de la consonne et de la voyelle subséquente est également propice à la bonne identification de la consonne (Bonneau *et al.*, 1996)

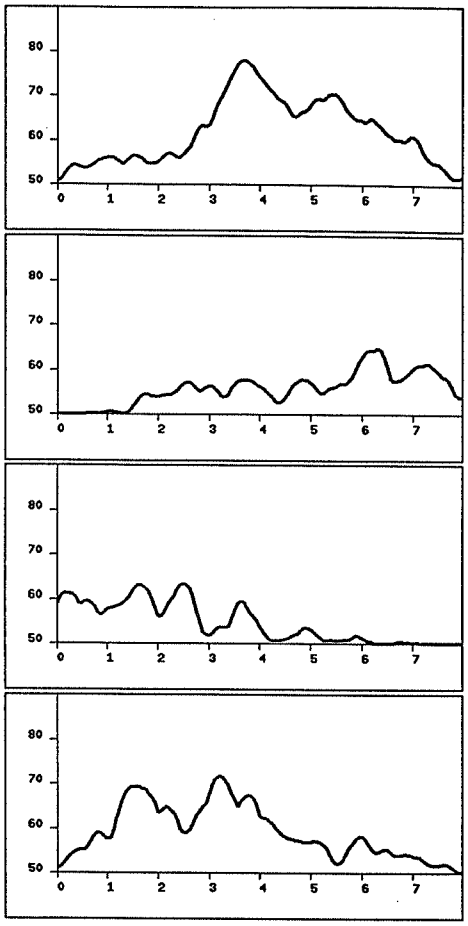


Figure 1: Spectres représentant le bruit de friction de l'occlusive, situé entre l'attaque (non incluse) et la voyelle. De haut en bas : /k/ suivie de /i/, /t/ suivie de /i/, /t/ suivie de /a/, /t/ suivie de /u/. La fréquence est indiquée en kHz sur l'axe des abcisses, et l'intensité en dB sur l'axe des ordonnées.

le sujet.

Le timbre vocalique semble, au moins en partie, déterminé par la répartition de l'énergie dans le spectre. En contexte /i/, un bruit de friction très intense et riche en fréquences élevées est généré au relâchement de l'articulation de /t/ et de /k/ (cf Fig. 1). Il est probable que la présence de ce bruit élevé favorise la perception de /i/, très bien identifié à partir de /t/ et de /k/. Le bruit est plus faible et moins élevé pour /p/ suivi de /i/, et de fait, le taux d'identification de cette voyelle est moins élevé (63%) quoique nettement supérieur à la chance. À l'opposé, plus l'énergie est concentrée dans les basses fréquences, plus le score d'identification de /u/ est élevé, comme en attestent les scores de /u/ sur /k/ (99%) et de /u/ sur /p/ (91%).

D'autres indices, en particulier le point de

départ des transitions, peuvent également contribuer à l'intelligibilité de la voyelle.

Nous avons étudié les relations entre les résultats de cette expérience et ceux de nos expériences précédentes (Bonneau *et al.*, 1996) concernant l'identification du bruit des occlusives et le rôle joué par la connaissance du contexte (l'identité de la voyelle subséquente était révélée aux auditeurs). Il n'y a pas de relation très claire entre l'identification globale de la voyelle et celle du bruit d'explosion. En effet, si toutes les données sont soumises à l'analyse, la corrélation entre les deux types d'identification n'est pas significative, et elle reste relativement faible ($r = 0.6$) si on écarte de l'analyse 20% des données les plus isolées. Si on considère les données contexte par contexte, en revanche, on observe quelques corrélations négatives, notamment pour les contextes /pi/ et /ki/, ainsi qu'une corrélation positive dans le contexte /ka/. Il semble que la présence d'un niveau d'énergie important dans les hautes fréquences favorise l'identification de /i/ et de /t/, ce qui pourrait expliquer les corrélations négatives entre l'identification de /i/ et celles de /p/ et de /k/. Quant à la corrélation positive entre les identifications de /k/ et de /a/, elle nous rappelle le principal résultat de notre expérience concernant l'influence de la connaissance préalable du contexte sur l'identification de l'occlusive : le taux d'identification de /k/ dans ce contexte avait augmenté de 18%. Rappelons également un résultat important (quoique attendu) qui ne peut apparaître dans notre étude sur la corrélation entre identification vocalique et consonantique : la proximité des articulations vocaliques et consonantiques entraîne toujours un taux d'identification élevé de la voyelle et de la consonne.

4. CONCLUSION

Notre expérience montre que les grands traits vocaliques peuvent être identifiés à partir du bruit d'explosion des occlusives françaises dépourvu de segment vocalique, puisque le score d'identification global atteint 80%. Le trait avant-arrière est particulièrement bien identifié puisque moins de 2% des voyelles /i/ et /u/ sont confondues l'une avec l'autre. Ces résultats suggèrent que, en l'absence même de tout segment vocalique, la voyelle joue probablement un rôle non-négligeable dans l'identification des occlusives.

5. BIBLIOGRAPHIE

Bonneau, A., Djezzar, L., & Laprie, Y. (1996). Perception of the place of articulation of French stop bursts. *J. Acoust. Soc. Am.* A paraitre.

Cullinan, W. L., & Tekieli, M. E. (1979). Perception of vowel features in temporally-segmented noise portions of stop consonant CV syllables. *J. Speech Hearing Res., 22*, 122–131.

Fant, G. (1969). Stops in CV-syllables. Tech. rep. 4, Royal Institute of Technology, Stockholm, Sweden. Also published in Speech Sounds and Features, The MIT Press, Cambridge , Massachussets, and London, England, 1973.

Kewley-Port, D., Pisoni, D. B., & Studdert-Kennedy, M. (1983). Perception of static and dynamic acoustic cues to place of articulation in initial stop consonants. *J. Acoust. Soc. Am., 73*(5), 1779–1793.

Krull, D. (1990). Relating acoustic properties to perceptual responses: a study of Swedish voiced stops. *J. Acoust. Soc. Am., 88*(6), 2557–2570.

Repp, B., & Lin, H. B. (1989). Acoustic properties and perception of stop consonant release transients. *J. Acoust. Soc. Am., 85*(1), 379–396.

Stevens, K. N., & Blumstein, S. E. (1978). Invariant cues for place of articulation in stop consonants. *J.Acoust. Soc. Am., 64*(5), 1353–1368.

Winitz, H., Scheib, M. E., & Reeds, J. A. (1972). Identification of stops and vowels from the burst portion of /p,t,k/ isolated from conversational speech. *J. Acoust. Soc. Am., 51*(4), 1309–1317.

INFLUENCE DU GENRE GRAMMATICAL DANS LA RECONNAISSANCE AUDITIVE DES MOTS EN FRANÇAIS

***Sophie MONPIOU, **Marie-Noëlle METZ-LUTZ, *François WIOLAND**

*Institut de Phonétique - Université des Sciences Humaines de Strasbourg
22, rue Descartes - 67084 Strasbourg Cedex
monpiou@ushs.u-strasbg.fr

**Hôpitaux Universitaires de Strasbourg - Clinique Neurologique
67091 Strasbourg Cedex

ABSTRACT

The aim of the study is to analyze auditory word recognition in French. More precisely, we propose to investigate, using lexical decision tasks and related paradigms, the possible role of the grammatical gender represented by the French singular article "le/la", both in mental lexicon access and in the organization of the mental lexicon.

INTRODUCTION

La compréhension d'un message verbal repose sur la capacité à reconnaître, c'est-à-dire à identifier, les constituants de ce dernier. Ainsi, la reconnaissance des mots, ou reconnaissance lexicale, constitue pour cette raison une étape essentielle dans le processus de compréhension du langage oral. Cette étape repose sur une comparaison entre la représentation de l'entrée verbale et les représentations mentales stockées dans le lexique interne (Segui, 1991). Celui-ci contient l'ensemble des représentations mentales des mots d'une langue donnée conservé dans la mémoire de tous les locuteurs de cette langue. Reconnaître un mot c'est, par conséquent, avoir accès à sa représentation mentale au sein du lexique interne.

L'étude de la reconnaissance lexicale soulève deux questions : la première concerne la façon d'accéder au lexique interne, la seconde a trait à l'organisation intrinsèque de celui-ci. D'après le modèle de la Cohorte (Marslen-Wilson & Welsh, 1978, Marslen-Wilson, 1987), la reconnaissance des mots procède à partir de deux types d'informations. D'une part des informations de type "bas-en-haut" ou informations acoustico-phonétiques fournies par le stimulus. D'autre part, des informations de type "haut-en-bas" qui sont des informations issues du contexte du stimulus. Le modèle de la Cohorte propose une description du processus de reconnaissance lexicale quelle que soit la langue dans laquelle il intervient. Or, les langues ne partagent pas toutes la même organisation structurelle. Par conséquent, les processus de reconnaissance lexicale ne dépendraient-ils pas, dans une certaine mesure, de l'organisation propre des langues? Selon Cutler & al. (1986), les comportements de sujets français et anglais diffèrent dans certaines situations comme par exemple la segmentation des unités verbales d'un message. Chaque langue élaborerait ainsi ses propres stratégies, ou ses propres "routines" dans le processus de reconnaissance lexicale.

Dans cette optique, nous avons décidé d'étudier le rôle du genre grammatical, représenté par l'article défini "le/la", dans les mécanismes d'accès au lexique interne en français.

Au sein d'une langue, le genre a pour objectif d'ordonner les éléments référentiels (c'est-à-dire les noms et les pronoms) en classes lexicales (Renault, 1987). Cette organisation particulière rend compte, de plus, des relations de dépendance qui existent entre différents constituants : en français, par exemple, à l'intérieur du SN le nom et l'adjectif doivent s'accorder en genre et en nombre (Lyons, 1972). Renault ajoute, de plus, que les classes lexicales possèdent chacune une marque particulière ou "trait de genre". Cette marque ne figure pas, en français, dans l'unité lexicale elle-même, mais dans un élément qui lui est associé : l'article singulier (en effet, au pluriel, le genre est neutralisé). Il faut noter que le genre est omniprésent en français. Comme le souligne Desrochers (1986), il n'existe aucun substantif qui ne possède le genre masculin ou féminin. Par conséquent, nous pouvons nous demander dans quelle mesure cet élément contextuel

exerce une influence sur l'accès au lexique dans le cadre du français.

EXPÉRIENCES

Expérience n°1 : Tâches de décision lexicale

Cette expérience se compose de deux conditions : la condition n°1 dans laquelle les stimuli (mots et non-mots) sont présentés isolément et la condition n°2 dans laquelle les stimuli (mots et non-mots) sont précédés de l'article défini singulier "le/la". D'après les résultats d'études antérieures (Huckel, 1991 ; Monpiou, 1995) la reconnaissance des mots précédés de leur marque de genre explicite devrait être plus rapide que celle des mots présentés hors de tout contexte.

Méthode

Sujets : 19 étudiants volontaires de langue maternelle française ont participé à cette expérience.
Matériel : une liste de 160 stimuli (80 mots et 80 non-mots) a été élaborée pour chaque condition. Dans chaque liste, les stimuli, séparés par 3 sec. de silence, étaient présentés dans un ordre aléatoire.
Procédure : Les sujets avaient pour consigne de déterminer le plus rapidement et avec le plus d'exactitude si les stimuli qu'ils percevaient étaient des mots ou non en appuyant soit sur un bouton "oui", soit sur un bouton "non". Les deux listes étaient présentées dans un ordre aléatoire aux sujets.

Résultats

Seuls le taux d'erreurs ainsi que les moyennes des temps de réaction (TR) concernant les réponses exactes ont été analysés. Les temps de réaction ont été mesurés à l'initiale de chaque séquence dans les deux conditions : le mot dans la condition n°1, l'article dans la condition n°2. Supposant que dans ces deux conditions le processus de reconnaissance du mot repose sur une analyse acoustico-phonétique mise en jeu dès que le sujet perçoit le premier indice de parole, la reconnaissance du mot dans le groupe nominal comme dans la condition isolée intervient dès lors que la quantité d'informations acoustico-phonétiques est suffisante pour une décision lexicale. Ainsi, en vue de comparer les TR liés directement à la reconnaissance du mot, dans les deux conditions, nous avons soustrait de la moyenne des TR des séquences "article + mot" et "article + non-mot", la durée moyenne de l'article ("le" = 137 ms ; "la" = 144 ms) en posant que la reconnaissance du genre de l'article intervient dans la durée physique de cet article. L'analyse de variance effectuée sur les moyennes des temps de réaction dans les deux conditions montre un effet significatif du facteur *condition* $F(1,128) = 16,992$ $p = .0001$. Ce résultat indique que lorsqu'un mot est précédé de sa marque de genre, c'est-à-dire lorsqu'il est précédé de l'article défini singulier, il est reconnu plus rapidement (TRm- article condition n°2 = 803 ms) que lorsqu'il est présenté isolément (TRm condition n°1 = 929 ms). Nous pouvons nous demander si cet effet est dû à l'article lui-même ou au genre qu'il porte. Pour répondre à cette question, nous avons élaboré une seconde expérience.

Expérience n°2 : Tâche de décision de genre

Le paradigme utilisé dans cette expérience est proche de celui de la décision lexicale : les sujets ont pour consigne de déterminer le genre grammatical des stimuli qu'ils perçoivent en appuyant soit sur un bouton "masculin" soit sur un bouton "féminin". Les 160 stimuli étaient uniquement des mots réels présentés isolément.

Méthode

La méthode utilisée est la même que celle de l'expérience n°1.

Résultats

Le pourcentage d'erreurs ainsi que la moyenne des temps de réaction concernant les réponses exactes ont été analysés. Nous avons comparé les TRm obtenus dans cette expérience (TRm = 927 ms) avec ceux concernant les mots en contexte de l'expérience n°1 (condition n°2). Pour pouvoir effectuer cette comparaison, la valeur moyenne de la durée de l'article a été soustraite des séquences "article + mots" de la condition n°2 de l'expérience n°1 (TRm - article = 803 ms). L'analyse de variance réalisée sur les moyennes des temps de réaction des deux types de stimuli montre un effet significatif du facteur *condition* $F(1,140) = 24,888$ $p = .0001$. Cela signifie que la tâche de décision lexicale portant sur les mots en contexte est significativement plus rapide que celle de détermination de genre. Ces résultats nous

permettent d'émettre l'hypothèse selon laquelle les représentations mentales seraient ordonnées dans le lexique interne d'après leur marque de genre explicite. En effet, d'après les résultats obtenus dans cette expérience, la décision de genre est post-lexicale c'est-à-dire qu'elle survient plus tardivement que la décision lexicale. Pour tenter de vérifier cette hypothèse, nous avons élaboré une troisième expérience.

Expérience n°3 : Tâche de décision de compatibilité

La consigne proposée aux sujets dans cette expérience est de déterminer si, dans les séquences "article + mot" qu'ils perçoivent, le genre de l'article est compatible avec celui du mot qu'il précède. Une liste de 160 stimuli "article + mot" a été réalisée : la moitié des séquences (80) étaient incompatibles en genre (ex. : *le maison), l'autre moitié était compatible (ex. : le courage).

Méthode

La méthode est la même que celle utilisée dans les deux expériences précédentes.

Résultats

Nous avons pris en compte pour l'analyse, le taux d'erreurs ainsi que la moyenne des temps de réaction des réponses exactes. Seuls les TRm des séquences "article + mot" de la condition n°2 de l'expérience n°1 (TRm = 943 ms) et ceux des séquences compatibles en genre de cette expérience (TRm = 1049 ms) ont été comparés. L'analyse de variance effectuée sur les temps de réaction montre un effet significatif du facteur *condition* $F(1,140) = 17,42$ $p = .0001$. Ainsi, la décision de compatibilité est significativement plus longue que la décision lexicale portant sur des mots en contexte.

DISCUSSION ET CONCLUSION

Les trois expériences précédemment décrites ont été réalisées dans le but d'étudier l'effet possible du genre grammatical à la fois dans le processus d'accès au lexique interne et dans l'organisation de ce dernier.

Les résultats obtenus dans la tâche de décision lexicale (exp. n°1) portant sur des stimuli isolés (condition n°1), nous permettent de déduire que le traitement d'un mot isolé n'est pas différent de celui d'un non-mot isolé. En effet, il n'y a aucune différence significative entre le temps mis pour identifier un mot hors contexte (TRm = 929 ms) et celui mis pour identifier un non-mot dans la même situation (TRm = 1016 ms). Ainsi, les sujets ont développé une stratégie différente de celle qu'ils utilisent dans la situation de communication normale. Dans la tâche de décision lexicale portant sur des stimuli isolés, par opposition à la communication verbale habituelle, il n'y a aucun indice contextuel susceptible de faciliter le processus de reconnaissance. Les sujets ont donc besoin d'un plus grand nombre d'informations acoustico-phonétiques pour pouvoir effectuer leur décision. Par conséquent, les mots ainsi que les non-mots isolés ont été reconnus respectivement bien après soit leur point d'unicité (Marslen-Wilson, 1984) soit leur point de déviation (Radeau & al., 1989). Analysés dans le cadre du modèle Cohorte II (Marslen-Wilson, 1987), ces résultats signifieraient que l'identification des stimuli ne pourrait avoir lieu qu'après un parfait appariement (dans le cas des mots) entre l'entrée sonore et sa représentation mentale.

Les résultats concernant les mots réels isolés et en contexte (exp. n°1) indiquent que lorsqu'un mot est précédé de sa marque de genre, celui-ci est reconnu significativement plus rapidement que lorsqu'il est présenté isolément (TRm condition n°1 = 929 ms ; TRm-article condition n°2 = 803 ms). Dans le modèle Cohorte II, les informations contextuelles (*top-down information*) posséderaient des "effets facilitateurs" sur le processus d'accès au lexique interne. Ces effets se traduiraient par une élévation du seuil d'activation de certains candidats par rapport à d'autres. Ainsi, nous pouvons supposer que le rôle de l'article défini, en tant qu'information contextuelle, serait d'élever le seuil d'activation des candidats de la cohorte initiale partageant le même genre que celui-ci. Cette activation sélective aurait pour conséquence de réduire le nombre de candidats potentiels et donc de réduire le temps nécessaire à la reconnaissance des stimuli. Dans le cas des stimuli isolés, il n'y a aucun indice contextuel permettant de réduire la taille de la cohorte initiale.

Les résultats de l'expérience n°2 indiquent que la décision de genre est post-lexicale c'est-à-dire qu'elle survient après la décision lexicale portant sur des mots en contexte. Ainsi, l'accès au genre d'un mot ne pourrait

se réaliser qu'après que l'accès au mot ait eu lieu. La stratégie utilisée par les sujets ne serait pas uniquement celle d'une analyse acoustico-phonétique du stimulus sonore mais également celle d'une recherche active dans le lexique. Une telle stratégie lexicale est induite par le type de tâche : les sujets savaient qu'ils n'avaient affaire qu'à des mots.

Les résultats de l'expérience n°3 indiquent que la tâche de décision de compatibilité est plus complexe que celle de reconnaissance des mots en contexte. D'après la consigne, les sujets savaient que le genre de l'article et celui du mot pouvaient, dans certains cas, ne pas être compatibles. Ainsi, l'effet du contexte, c'est-à-dire de l'article, a été neutralisé par la situation expérimentale. Aucune des représentations mentales de la cohorte initiale n'avait un seuil d'activation plus élevé qu'une autre en fonction de son genre, puisque l'article ne pouvait constituer une information valable. L'effet du genre n'était plus alors absolu, comme dans la tâche de décision lexicale, mais il était devenu plus "probalistic" (Grosjean & al., 1993).

D'après les résultats obtenus aux trois expériences, le genre grammatical jouerait un certain rôle dans l'organisation du lexique interne en français. Nous pouvons supposer que celui-ci ressemblerait à un dictionnaire uniquement composé de mots accompagnés de leur marque de genre.

Actuellement, d'autres expériences sont en cours pour tenter de disposer de données psychophysiologiques (mesures de potentiels évoqués) dans des tâches similaires à celles décrites précédemment. Le but de ces données est de tenter d'établir des corrélations entre les données comportementales et les données psychophysiologiques pour préciser le rôle du genre grammatical dans le processus d'accès au lexique en français.

RÉFÉRENCES

Cutler A., Mehler J., Norris D., Segui J. (1986) The syllable's differing role in the segmentation of French and English. *Journal of Memory and Language*, n° 25, 385-400.

Desrochers A. (1986) Genre grammatical et classification nominale. *Revue Canadienne de Psychologie*, n° 40 (3), 224-250.

Grosjean, F., Dommergues, J.Y., Cornu, E. (1992) The gender marking effect in spoken word recognition. *Perception & Psychophysics*, n° 56 (6), 590-598.

Huckel C. (1991) De l'importance du genre dans la décision lexicale. Mémoire de DEA dirigé par F. Wioland, Laboratoire de Phonétique de Strasbourg, USHS.

Lyons J. (1972). *Linguistique générale. Introduction à la linguistique théorique.* Coll. "Langue et Langage", Larousse, Paris.

Marslen-Wilson W., Welsh A. (1978) Processing interactions and lexical access during word-recognition in continuous speech. *Cognitive Psychology*, n° 10, 29-63.

Marslen-Wilson W. (1984) Function and process in spoken word recognition. A tutorial review. In H. Bouma and D.G. Bouwhuis (Eds.) *Attention and Performance X : Control of Language Processes*, Hillsdale, NJ : L. Erlbaum, 125-150.

Marslen-Wilson W. (1987). Functionnal parallelism in spoken word recognition. *Cognition*, n° 25, 71-102.

Monpiou S., Metz-Lutz M-N., Wioland F. (1995) La reconnaissance auditive des mots en français. Rôle du genre grammatical porté par l'article défini. *TIPS*, n° 25, 21-47.

Monpiou S., Metz-Lutz M-N., Wioland F. (1995) Role of the grammatical gender in mental lexicon access. *ICPhS*, Stockholm, 13-19 August 1995, vol. 3, 596-599.

Radeau, M., Morais, J. (1989) The effect of the uniqueness point in shadowing spoken words. *Speech Communication*, n° 9, 155-164.

Segui J. (1991) La reconnaissance visuelle des mots. In R. Kolinsky, J. Morais, J. Segui (Eds.) *La reconnaissance lexicale dans les différentes modalités sensorielles : Études de Psycholinguistique Cognitive.* P.U.F., 100-117.

Renault R. (1987) Genre grammatical et typologie linguistique. *Bulletin de la Société de Linguistique de Paris*, 82 (1), 69-117.

LA SEGMENTATION LEXICALE: CONTRIBUTION DES INDICES SEGMENTAUX ET SUPRASEGMENTAUX

Christine Meunier[1], Uli Frauenfelder[1] et Alain Content[1,2]

[1] Laboratoire de Psycholinguistique Expérimentale,
FAPSE - Universite de Geneve - 9 route de Drize - CH-1227 Carouge - SUISSE
tél: (41) 22 705 97 41 - e-mail: meunier@fapse.unige.ch

[2] Laboratoire de Psychologie Expérimentale, Université libre de Bruxelles

ABSTRACT

Previous studies have shown the importance of prosodic features for lexical parsing. More precisely, the importance of rhythm was demonstrated in the segmentation of ambiguous bisyllabic words. This paper evaluates the weight of different features (duration, intonation, coarticulation, intensity) in lexical parsing. Our results show that temporal modifications, as well as other features, induce changes in word segmentation.

1. INTRODUCTION

Un des problèmes majeurs de la psycholinguistique est la compréhension des mécanismes qui entrent en jeu dans la reconnaissance des mots de la chaîne parlée. Si le code orthographique permet aisément d'isoler les mots dans une phrase, le signal acoustique ne laisse pas apparaître d'indices aussi évidents que les "blancs" séparant les mots sur une page de texte. Ainsi, le problème de la reconnaissance et donc l'extraction d'unités dans un flux continu nous renvoie directement au problème de la segmentation.

L'analyse acoustique permet de mettre en évidence plusieurs types d'indices relevant de niveaux d'analyse différents: d'une part les indices segmentaux (l'information allophonique et la coarticulation) et d'autre part, les indices suprasegmentaux ou prosodiques (durée, intonation et amplitude du signal).

L'utilisation des indices prosodiques par les auditeurs a clairement été démontrée pour l'anglais (Cutler et Norris, 1988). En français, Banel et Bacri (1994) ont montré que la segmentation lexicale est influencée par les indices rythmiques. Ces auteurs ont utilisé des mots bisyllabiques composés de deux monosyllabes enchassés. Les sujets devaient décider si l'expression présentée correspondait à deux mots ou à un seul. En manipulant les durées relatives des deux voyelles, Banel et Bacri ont constaté que le rythme iambique (bref-long) favorisait le groupement des deux syllabes (réponse 1 mot) tandis que le rythme trochaïque (long-bref) augmentait leur séparation. Cependant, dans cette étude, les stimuli étaient toujours réalisés à partir de la concaténation des deux monosyllabes. On peut donc penser que les indices acoustiques non manipulés favori-saient la séparation. Toutefois, d'autres études ont montré l'effet de la durée en tenant compte des autres facteurs (Rietveld, 1980).

Notre objectif est d'évaluer la contribution relative des différents types d'indices (rythme d'une part, coarticulation, intonation et amplitude, d'autre part). Pour ce faire, nous avons modifié l'organisation temporelle de mots bisyllabiques, en partant soit du mot bisyllabique, soit d'une concaténation de leur composants monosyllabiques. Nous obtenons ainsi des stimuli pour lesquels les indices de durée entrent en conflit avec l'information induite par les autres indices (coarticulation, intonation et amplitude du signal). Deux expériences ont été menées, en variant la tâche proposée aux sujets.

2. PLAN EXPERIMENTAL

Quatre versions des stimuli ont été préparées afin de mettre en compétition différents types d'indices.

La première version (V1) est constituée par l'enregistrement naturel du mot bisyllabique (exemple: "corsage"). La deuxième version (V2) est constituée des deux mots monosyllabiques concaténés ("corps", "sage"). Les versions 3 et 4 sont obtenues en appliquant le schéma rythmique de V1 sur V2 et inversement (table 1):

Table 1: définition et élaboration des versions selon leur schéma rythmique et leur enchaînement segmental: en gras les versions "d'origine", en normal, les versions "modifiées".

	Bref-Long	Long-Long
non concaténés	V1 -------> V4	
concaténés	V3 <------- V2	

Ainsi, pour les versions 1 et 2 les indices phonétiques dont dispose l'auditeur sont congruents (table 2). A l'inverse, pour les versions 3 et 4, l'information temporelle est en conflit avec l'information apportée par les autres indices phonétiques. La compétition

entre les différents types d'indices devrait donc entraîner une certaine ambiguïté dans le traitement de ces versions.

Table 2: type d'information (en terme de segmentation lexicale: 1 mot ou 2 mots) apportée par les différents indices pour les 4 versions

	V1	V2	V3	V4
durée	1	2	1	2
coarticulation	1	2	2	1
intonation	1	2	2	1
amplitude	1	2	2	1

La comparaison entre ces quatre versions nous permet de dégager la prédominance éventuelle de certains types d'indices pour la segmentation lexicale. Deux hypothèses s'offrent à nous:

1/ Il existe une dominance d'un type d'indice sur les autres et dans ce cas, soit: a) les indices temporels sont prédominants. V3 serait alors interprétée majoritairement comme 1 mot bisyllabique et V4 comme 2 mots monosyllabiques; b) les autres indices sont prédominants, et V3 serait interprétée comme 2 mots et V4 comme 1 mot.

2/ il n'y a pas de dominance entre les indices et dans ce cas l'interprétation de V3 et V4 est ambiguë.

3. EXPERIENCE I

3.1. Matériel linguistique

Nous avons sélectionné 24 mots bisyllabiques de la langue française dont chaque syllabe isolée forme un mot monosyllabique (ex. "bulgare" / "bulle" "gare").

Les 24 mots bisyllabiques ainsi que les 48 mots monosyllabiques ont été enregistrés par un locuteur. La durée de chaque syllabe des mots monosyllabiques et des bisyllabiques a ensuite été mesurée. Ces mesures mettent en évidence une réduction notable de la durée de la première syllabe des bisyllabes (table 3). Ces observations sont conformes aux remarques de Rossi et al. (1981) qui montraient un fort allongement de la syllabe finale par rapport aux syllabes précédentes, à l'intérieur d'un groupe rythmique.

Table 3: durées moyennes (en ms) des syllabes S1 et S2 dans les mots bisyllabiques et monosyllabiques.

	bisyll. *(ex. bulgare)*	monosyll. *(ex. bulle gare)*
S1	233	453
S2	366	359

Les versions 1 et 2 constituent les versions "naturelles". Les versions 3 et 4 sont obtenues à l'aide d'un calcul mettant en évidence les relations temporelles liant S1 et S2 dans les couples de monosyllabes et dans les mots bisyllabiques. Ce sont les versions "modifiées".

Version 3

La version 3 est obtenue en appliquant le schéma temporel du mot bisyllabique ("bulgare") sur la suite de monosyllabes ("bulle"+"gare"): la durée du mot monosyllabique correspondant à S1 est réduite de façon à ce que le rapport S1/S2 soit comparable à celui que l'on trouve dans le mot bisyllabique.

Cette formule est appliquée à chaque item de façon à ce que le rapport S1/S2 soit spécifique à chaque mot. Cette précaution nous semble nécessaire dans la mesure où les rapports S1/S2 sont assez variables d'un item à l'autre en fonction des unités phonétiques présentes.

Version 4

La relation inverse est appliquée de façon à obtenir la version 4 à partir d'une transformation de V1. Dans V4, le mot "bulgare" reçoit ainsi un schéma temporel comparable à celui de "bulle"+"gare" dans V2, par allongement de la 1ère syllabe.

L'allongement et la compression ont été réalisés à l'aide du logiciel SoundDesigner II.

3.2. Procédure

Les stimuli ont été enregistrés sur cassette. Chaque stimulus était présenté 3 fois avec un intervalle d'une seconde.

L'ensemble des 96 stimuli ont été répartis en 4 listes, chacune contenant l'ensemble des 24 stimuli. Les 4 versions de chaque item ont été réparties de manière équilibrée entre les listes.

Les sujets (n=49) ont été répartis en 4 groupes. Chaque groupe a entendu une des 4 listes, précédée de 4 exemples.

Les sujets devaient transcrire ce qu'ils entendaient et avaient la possibilité de proposer plusieurs transcriptions en les classant par ordre de préférence. Le temps de réponse des sujets était limité à 10 secondes.

3.3. Résultats

Les résultats de cette première expérience montrent que pour les deux versions non modifiées (V1 et V2) les réponses des sujets expriment clairement l'information apportée par les deux types d'indices (table 2): V1 est interprétée comme 1 mot bisyllabique et V2

majoritairement comme 2 mots monosyllabiques (figure 1).

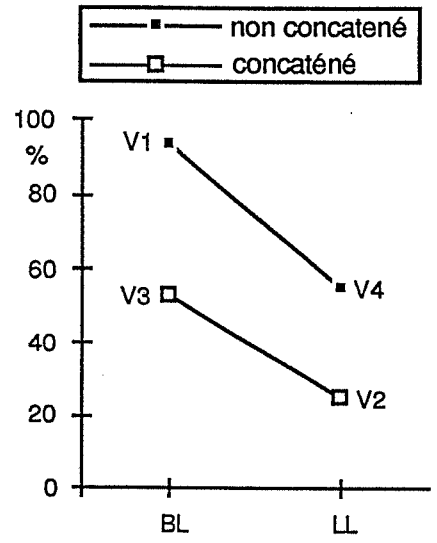


Figure 1: Taux de réponses "un mot".

Pour les deux versions modifiées (V3 et V4), les résultats sont moins contrastés: pour chacune des versions on observe autour de 50% de réponses "1 mot".

La modification du rythme dans un mot bisyllabique (V1->V4) réduit de 40% le taux de groupement ($p<.0001$), tandis que la modification rythmique inverse dans deux mots monosyllabiques (V2->V3) réduit de 30% le taux de séparation ($p<.0001$). Par ailleurs, la modification des autres indices (coarticulation, intonation, intensité) entraîne également des différences significatives—pour une même structure rythmique (V1->V3, $p<.0001$; V2->V4, $p<.0001$).

La tâche de choix libre donne lieu à 6.8% de réponses "non attendues", c'est-à-dire celles qui ne correspondent ni au groupement ni à la séparation (V1: 3.5%, V2: 4.6%, V3: 9.6%, V4: 9.6%). On observe donc, dans l'ensemble, que la majeure partie des réponses se répartissent entre les deux catégories attendues.

Au vu de cette première expérience, il semble que la deuxième hypothèse exposée plus haut soit observée: il n'apparaît pas de dominance des différents types d'indices, puisque la manipulation du rythme montre un effet aussi bien pour la version concaténée que pour la version non-concaténée.

Une deuxième expérience a été mise au point afin d'approfondir et de préciser les résultats de l'expérience 1. La tâche de choix forcé devrait éliminer les réponses non attendues et donc augmenter le pourcentage de réponses 1 mot pour V1 et de réponses 2 mot pour V4. Par ailleurs, la mesure des temps de réaction (TR) devrait nous permettre de confirmer la difficulté de traitement des stimuli ambigus.

4. EXPÉRIENCE 2

4.1. Matériel linguistique

Les stimuli de l'expérience 2 sont identiques à ceux de l'expérience 1.

4.2. Procédure

L'ensemble des 25 sujets ont entendu la totalité des stimuli. Les 4 listes, et les items dans chaque liste, étaient présentées dans un ordre aléatoire.

Les sujets avaient une présentation visuelle des deux possibilités (1 mot bisyllabique ou 2 mots monosyllabiques) et devaient, après audition du stimulus, choisir l'une des alternatives en appuyant sur l'un des deux boutons d'un boîtier de réponse le plus rapidement possible. L'expérience a été réalisée à l'aide du logiciel Psyscope.

4.3. Résultats

Les résultats confirment et renforcent ceux obtenus pour l'expérience 1. Cette fois, les sujets sont unanimes à interpréter la version "mot bisyllabique" comme 1 seul mot, et la version "mots concaténés" comme une suite de deux mots monosyllabiques (figure 2). La tâche de choix forcé a donc contrasté les réponses concernant les deux versions naturelles.

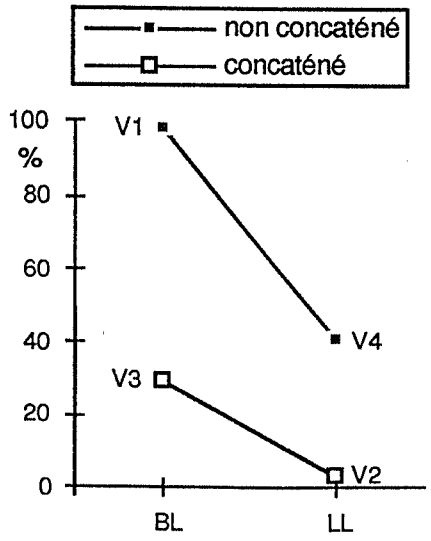


Figure 2: Taux de réponse "un mot"

Les résultats pour les deux manipulations du rythme donnent lieu à des taux de réponses intermédiaires. Les deux manipula-tions opposées ont des effets systématiques et statistiquement significatifs. L'allon-gement de la première syllabe dans les mots bisyllabiques réduit de plus de 50% le taux d'interprétations

"mot bisyllabique" (p<.0001). Inversement, un rythme bref-long sur les paires de mots concaténés augmente de 25% l'interprétations "mot bisyllabique" (p<.0001). Les variations de réponses dues aux facteurs communs (coarticulation, intonation, intensité) sont également significatives. Enfin, les résultats montrent une interaction (p<.0002), puisque l'effet de la manipulation du rythme est plus marqué sur la version non-concaténée, qui préserve les indices de coarticulation, d'intonation et d'amplitude. Ce résultat confirme donc l'importance de l'indice de durée pour la segmentation lexicale.

Toutefois, il faut noter une très grande variabilité des réponses pour les versions modifiées. Cette variabilité semble être fonction des stimuli.

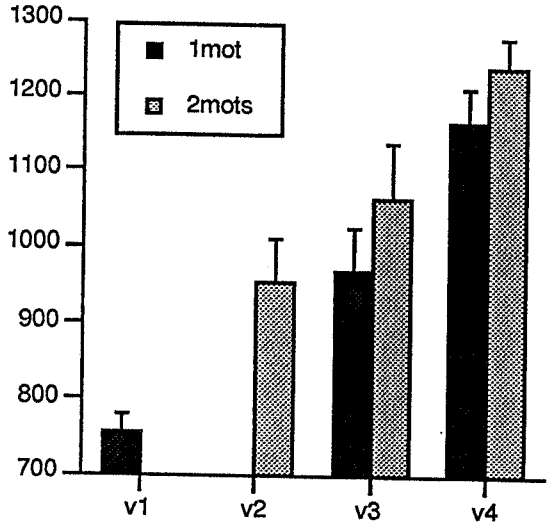


Figure 3: temps de réaction moyen (en ms) des réponses "1 mot" et "2 mots".

Les TR observés sont systématiquement fonction du degré d'ambiguïté du stimulus. Les temps les plus courts sont observés dans les deux conditions où les sujets montrent la plus grande unanimité (figure 3, V1: 755ms; V2: 953ms). Entre les deux conditions intermédiaires, les temps sont plus longs dans la situation bisyllabes long-long (V4), qui correspond au taux de réponses le plus proche de 50%.

Ces observations nous permettent de confirmer la difficulté de traitement des stimuli ambigus.

5. DISCUSSION ET CONCLUSION

Les conditions énumérées au début de l'article permettaient d'envisager deux hypothèses: soit les informations apportées par les différents indices étaient traitées différemment selon le poids des indices, soit elles ne l'étaient pas.

L'hypothèse d'indices dominants n'a pas été confirmée. Il semble effectivement que les informations apportées par les indices temporels, d'une part, et celles apportées par les autres indices, d'autre part, ne permettent pas un type de réponse majoritaire lorsqu'elles sont en contradiction.

Ainsi, lorsque toutes les informations sont congruentes, le traitement est simple, lorsqu'elles sont en conflit, l'interprétation du signal devient difficile et contradictoire.

Il nous faut toutefois nuancer ces propos en évoquant la très forte variabilité des réponses en fonction des items. Il semble possible que la conjonction de certains facteurs favorise le poids d'un indice. Par ailleurs, une manipulation plus approfondie et plus fine des stimuli pourrait sans doute nous permettre de mieux comprendre cette forte variabilité. En particulier, il pourrait être intéressant de tenter de dissocier expérimentalement les effets des indices d'intonation, d'amplitude et de coarticulation que nous n'avons pas distingués dans cette première exploration.

L'intérêt de ces expériences est de mettre en opposition différents types d'indices phonétiques afin d'évaluer leur contribution pour la segmentation lexicale. Nous pensons par la suite poursuivre ces expériences en utilisant des tâches moins directement liées à la segmentation (comme une tâche de décision lexicale, par exemple).

6. BIBLIOGRAPHIE

Banel M.H., Bacri N. (1994) "On metrical patterns and lexical parsing in French", *Speech Communication*, 15, pp. 115-126.

Cutler A., Norris D. (1988) "The role of strong syllables in segmentation for lexical access", *J. Experimental Psychology: Human Perception and Performance*, Vol. 14, pp. 113-121.

Rietveld A.C.M. (1980) "Word boundaries in the French language", *Language and Speech*, Vol. 23, No. 3, pp. 289-296.

Rossi M., Di Cristo A., Hirst D.J., Martin P., Nishinuma Y. (1981) *L'intonation: de l'acoustique à la sémantique*. Paris, Klincksieck.

Remerciements

Cette recherche a été financée par le F.N.R.S. 11-39553.93. Nous remercions Marie Bachmann et Isabelle Gunther pour leur aide à la réalisation de l'expérience 1.

LE RÔLE DE LA COARTICULATION DANS LA PERCEPTION DES VOYELLES DE L'ARABE STANDARD MODERNE

Imad ZNAGUI, Mohamed YEOU

Laboratoire de Phonétique de l'Université Paris III, URA 1027, 19 Rue des Bernardins 75005, Paris,
Département de Linguistique arabe de l'Université Paris VIII

ABSTRACT

In Znagui (1995), the investigation of the coarticulatory influence of lingual consonants differing in place of coarticulation (interdental /ð, θ/, alveolar /s, z/, palatal /ʃ, ʒ/, postpalatal /k/, uvular /χ, ʁ, q/, pharyngeal /ħ, ʕ/ and pharyngealized /tˤ, sˤ, dˤ, ðˤ/ on the adjacent vowels /a, a:, i:, i:, u, u:/ in Modern Standard Arabic (MSA) was made. Measurements of the distance between the frequency of F1 and F2 in vowel steady were done. The results show that two categories of vowels can be distinguished as function of the distance between F1 and F2. The objective of this study is to investigate the perceptual significance of this acoustico-phonetic classification. One question of interest is to see if native speakers of Arabic can discriminate between these categories of vowels.

1. INTRODUCTION

Badreddine (1977) et Ghazeli (1977) ont observé que le déplacement horizontal de la langue en avant et en arrière du conduit vocal dans la production des consonnes de l'arabe antérieures et centrales (interdentales /ð, θ/, alvéolaires /s, z/, palatales /ʃ, ʒ/, postpalatal /k/), postérieures (uvulaires /χ, ʁ, q/, pharyngales (/ħ, ʕ/) et pharyngalisées dorénavant emphatiques (/tˤ, sˤ, dˤ, ðˤ/) a des conséquences sur l'articulation et les structures formantiques des voyelles adjacentes. Mais ces observations n'ont pas été vérifiées par des analyses acoustico-perceptives. D'ailleurs, les études acoustiques antérieures sur les réalisations des voyelles des dialectes arabes ou de l'Arabe Standard Moderne (ASM) en contexte ont montré une grande dispersion acoustique relative aux trois timbres phonologiques /a, i, u/ (Rajouani *et al.*, 1987; Belkadi, 1984; Kiel, 1987; Metoui, 1989). Mais ces auteurs raisonnent en termes de valeurs absolues relatives aux deux premiers formants (F1 et F2) des voyelles sans pour autant justifier leurs analyses au niveau perceptif. Ce qui nous amène à dire que l'espace vocalique de l'ASM et des dialectes arabes n'a pas été abordé de manière exhaustive mettant en jeu un cadre théorique précis et une démarche méthodologique rigoureuse et reproductible. De ce fait, nous nous intéressons ici à l'effet de la coarticulation gauche-droite sur la perception des voyelles de l'ASM en structure CV nécessaire à la synthèse par règles de cette langue.

2. OBJECTIF

L'objectif de ce travail est de tester la validité d'une classification acoustico-phonétique des voyelles de l'ASM au contact des consonnes antérieures (groupe1) , postérieures et emphatiques (groupe2) au niveau perceptif.

La première question est d'examiner si les auditeurs arabophones discriminent les deux catégories vocaliques (G1) et (G2).

La seconde question concerne le degré de coarticulation en structure CV: dans quel contexte consonantique la voyelle change de timbre ?

Les résultats de l'expérience perceptive seront discutés à la lumière des deux théories antagonistes : la théorie motrice (Liberman & Mattingly, 1985) et la théorie auditive (Chistovich & Lublinskaya, 1979; Schwartz, 1987, entre autres).

3. METHODE

3.1. Choix du corpus

90 stimuli ont été construits à partir d'un corpus de douze mots et logatomes produits par un locuteur marocain. Afin de conserver une certaine homogénéité dans la construction de ce corpus, certains critères ont été élaborés:

- nous n'avons retenu que les mots bisyllabiques de type: CV:CV et CV:CVC.

- les voyelles /a:, i:, u:/ étudiées apparaissent en syllabes ouvertes longues CV:

- l'entourage consonantique a été limité aux quatre contextes consonantiques : alvéolaire /s/, pharyngale /ʕ/, uvulaire /χ/ et emphatique /sˤ/.

- le contexte consonantique alvéolaire /s/ au contact de /a:, i:, u:/ sera pris comme référence de la classe des consonnes antérieures et centrales du fait que les voyelles précédées de celles-ci présentent une grande homogénéité acoustique par rapport aux voyelles précédées des consonnes postérieures et emphatiques Znagui (1995).

- nous avons retenu pour l'étude les voyelles accentuées longues appartenant à la première syllabe longue, dans la mesure où nos études acoustiques antérieures ont montré que ni la durée, ni l'accent ne jouent un rôle déterminant sur la qualité et la perception des voyelles de l'ASM Znagui (1995). De même, nous voulons confirmer l'hypothèse que les consonnes emphatiques n'influent pas uniquement la qualité des voyelles brèves inaccentuées /i, a, u/, mais également les voyelles /a':, i':, u':/fortes par leur durée longue et l'accent.

- dans le corpus, la consonne qui suit la voyelle longue accentuée est généralement une consonne labiale pour éviter l'effet de coarticulation.

- nous avons été amené à utiliser des logatomes dans certains contextes pour éviter la présence des consonnes nasales et postérieures.

Nous présentons un exemple de comparaisons de stimuli vocaliques présenté à nos sujets :

[a:] de [sa:]+ [a:] de [sˤa:]

[a:] de [sa:]+ [a:] de [χa:]

[a:] de [sa:]+[a:] de [ʕa:]

3.2. Matériel technique et procédure expérimentale

L'enregistrement du corpus a été effectué sur une bande magnétique en chambre insonorisée du Laboratoire de Phonétique de l'Université Paris III. On a utilisé à cet égard un magnétophone Revox A71, tournant à la vitesse de 7,5 cm/s. La bouche du locuteur était à une distance d'environ 25 cm du microphone (ECM 265). Les mots lus ont été écrits en arabe sur des feuilles séparées, présentées au fur et à mesure au locuteur. Chaque mot a été lu trois fois avec une voix normale. Seule une réalisation a été retenue pour l'expérience après l'avoir visualisée en spectrogramme à bande large à l'aide du programme UNICE. Nous avons donc pris la portion du signal où l'allure de la fréquence fondamentale est monotone. Les stimuli vocaliques choisis ont été digitalisés et édités par le programme d'acquisition et de traitement du signal Audiomédia sur Macintosh II. On a procédé par une technique de découpage de la partie stable de la voyelle en évitant les transitions, dans la mesure où notre objectif est de tester si l'effet des consonnes sur les voyelles dépasse la transition et modifie la cible acoustico-perceptive. La durée de chaque stimulus est de 300ms. Pour obtenir cette durée, nous avons dilaté le signal à l'aide du programme Audiomedia pour dupliquer des portions du signal (périodes) en prenant en considération le passage par zéro afin d'éviter les coupures abruptes du stimulus. Les stimuli ont été jugés de qualité satisfaisante par 5 auditeurs arabes. Par la suite, on a formé 3 blocs comprenant chacun 30 stimuli : le premier bloc contient les stimuli de [a:], le second de la voyelle [i:] et le troisième ceux de [u:]. Ces trois blocs ont été soumis à 10 auditeurs arabophones maghrébins (masculins) et 10 auditeurs francophones âgés entre 30 à 45 ans: ceux-ci ne présentent pas de troubles auditifs connus.

Avant la présentation auditive de l'expérience à l'aide du magnétophone Revox A71 liés à des hauts parleurs, nous avons familiarisé les sujets avec la tâche à l'aide d'un bloc d'entraînement. Nous avons choisi la tâche **AX** qui est une tâche moins coûteuse. Dans ce paradigme, les sujets doivent choisir entre deux réponses : identiques ou différentes (1 ou 2).

Paradigme /Stimuli / Réponse correspondante

AX fixe (AA) 1= identique
(AB) 2= différent

La passation individuelle a eu lieu dans la chambre insonorisée du laboratoire de Phonétique l'Université Paris III. Les stimuli sont présentés à un niveau d'intensité jugé confortable par les sujets. L'intervalle entre 2 stimuli qui compose la paire (AX) est de 200 ms, de 2 s entre chacune des paires et de 10 s entre deux blocs. Les sujets ont pour tâche de décider après chaque paire : A est différent de X ou identique et de noter leurs réponses sur papier (= ou #). Le test a duré environ 20 mn.

4. RÉSULTATS

La figure 1 indique que les 10 auditeurs arabophones sont capables de discriminer deux timbres vocaliques pour /a:, i:, u:/ dans la comparaison alvéolaire/emphatique

(alv-emp) mais non dans les comparaisons alvéolaire/uvulaire (alv-uvul) (19%, 17%, 22%) et alvéolaire/pharyngal (alv-phar) (13%, 20%, 16%). Une ANOVA à mesures répétées par sujets montre que le taux de discrimination diffère de manière significative entre les trois types de comparaisons précitées [$F_1(2,9)=46$, $p<0.0001$] pour /a:/, [$F_1(2,9)=45$, $p<0.0001$] pour /i:/ et [$F_1(2,9)=34$, $p<0.0001$] pour /u:/. L'analyse post-hoc révèle qu'il n'y a pas de différences significatives entre les deux comparaisons alv-uvul et alv-phar [Scheffé F-test, $p=0.05$] pour /a:, i:, u:/.

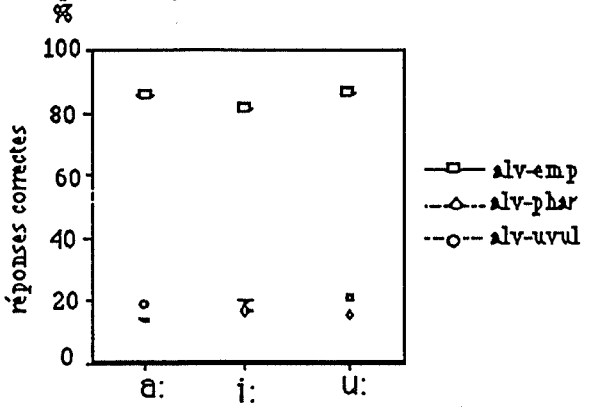


Figure 1 : Taux de discrimination des trois voyelles dans les trois comparaisons (alv-emp, alv-phar, alv-uvul) calculé sur les 10 auditeurs arabophones.

Ce résultat semble surprenant, dans la mesure où les timbres vocaliques discriminés dans la comparaison alv-emp pour /i:/, /a:/ et /u:/ n'ont pas de statut phonologique en ASM qui possède 3 voyelles brèves en opposition de quantité /a, a:, i, i:, u, u:/. A ce titre, nous pensons que la capacité de discrimination de nos sujets a des explications psychoacoustiques en se référant au rôle que peut jouer particulièrement la distance entre les deux premiers formant (rapprochement et éloignement de F1 et F2) dans la perception des voyelles Amerman & Daniloff (1977).. Dans le but d'examiner cette hypothèse, nous avons effectué une expérience de contrôle qui a conservé la même procédure que l'expérience précédente soumises à 10 auditeurs francophones qui possèdent un système phonologique vocalique riche de voyelles intermédiaires (11 ou 12 phonèmes). Les résultats montrent que les taux de discrimination dans la comparaison alv-emp sont de l'ordre de 98%, 87%, et 100% pour /a:, i:, u:/ respectivement. et 27%, 24%, 15% dans la comparaison alv-phar, et 28%, 26%, 20% dans la comparaison alv-uvul.

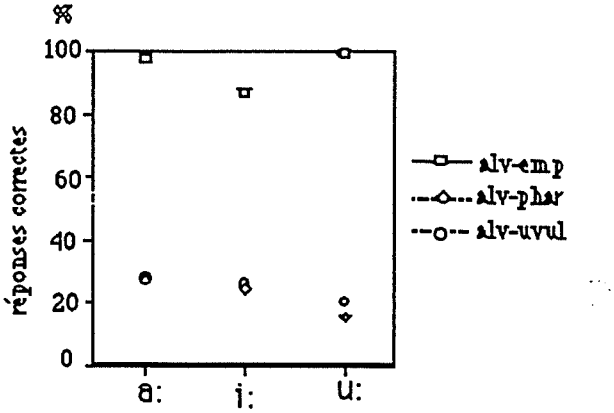


Figure 2 : Taux de discrimination des trois voyelles dans les trois comparaisons (alv-emp, alv-phar, alv-uvul) calculé sur les 10 auditeurs francophones.

Une ANOVA à mesures répétées par sujets révèle que le taux de discrimination diffère de manière significative entre les trois types de comparaisons [$F_1(2,9)=7$, $p<0.0001$] pour /a:/, [$F_1(2,9)=76$, $p<0.0001$] pour /i:/ et [$F_1(2,9)=207$, $p<0.0001$] pour /u:/. L'analyse post-hoc indique qu'il n'y a pas de différences significatives entre les deux comparaisons alv-uvul et alv-phar [Scheffé F-test, $p=0.05$] pour /a:, i:, u:/. La comparaison entre les performances des 10 auditeurs arabophones et des 10 auditeurs francophones montre une différence significative [$t(9)=4.3$, $p<0.0001$]. Ce qui suggère que les francophones ont une meilleure discrimination dans la comparaison alv-em par rapport aux arabophones grâce à leur système linguistique riche en voyelles. Toutefois, l'hypothèse de la prédominance du facteur psychoacoustique sur l'influence du système linguistique dans la tâche de discrimination des timbres vocaliques de l'ASM est confirmée par le fait que nos sujets arabophones ou francophones ont discriminé les deux timbres pour chacune des voyelles /a:, i:, u:/ dans la comparaison alv-emp.

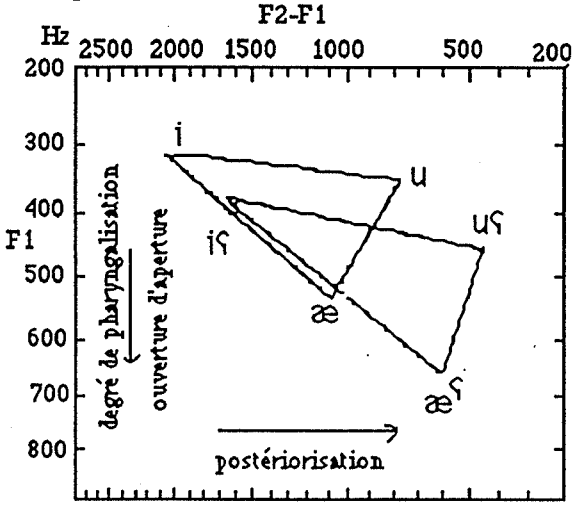
5. DISCUSSION

Le degré de rapprochement des voyelles /a:, i:, u:/ précédée des consonnes postérieures (pharyngale et uvulaire) ne semble pas être suffisant pour être audible chez nos auditeurs arabes et permettre par conséquent la discrimination de celle-ci par rapport à la voyelle précédée des consonnes antérieures. Cela suggère que le déplacement horizontal du corps de la langue en avant et arrière et en avant du conduit vocal dans la production des consonnes de l'ASM n'a pas d'effet audible sur la perception des voyelles adjacentes. Le trait antérieur/postérieur ne peut être considéré comme un trait distinctif pour la description phonologique des voyelles de cette langue. Il

reste que seules les consonnes emphatiques réalisées avec une double constriction alvéolaire et pharyngalisée arrivent à changer la cible acoustico-perceptive des voyelles adjacentes qu'elles soient longues ou brèves fermées comme /i, i:, u, u:/ ou ouvertes /a, a:/.

Une classification de nature psychoacoustique en deux catégories de voyelles emphatisées et non emphatisées de l'ASM serait pertinente, basée sur le facteur de rapprochement et d'éloignement entre les deux premiers formants (figure 3) :

- une catégorie de voyelles brèves et longues au contact des consonnes emphatiques reconnue par un rapprochement des deux premiers formants,
- une catégorie de voyelles brèves et longues au contact des consonnes non emphatiques reconnue par l'éloignement des deux premiers formants.



Figue 3 : Trapèze vocalique des voyelles brèves emphatisées et non emphatisées de l'ASM produites par 6 locuteurs maghrébins (142 réalisations par voyelle)

D'après les résultats obtenus, nous constatons qu'une classification acoustico-phonétique binaire en deux groupes de voyelles précédées des consonnes antérieures (G1), postérieures et emphatiques (G2) n'est pas forcément validée par la perception. Il existe donc une certaine divergence entre la production et la perception des voyelles de l'ASM. Il est utile de signaler que les locuteurs arabophones "bénéficient" d'une grande latitude articulatoire dans la production des trois timbres /a:, i:, u:/ et qui est corrélée à des dispersions acoustiques importantes dues au contact des consonnes de 8 différents lieux d'articulation (labial, interdental, alvéolaire, palatal, uvulaire, pharyngal, pharyngalisé et glottal) de l'ASM. Mais nos résultats montrent que cet espace vocalique est divisé en deux catégories au niveau psycho-acoustique par deux contextes consonantiques adjacents: emphatique et non emphatique. Sur le plan phonologique, nous pouvons postuler que chaque phonème vocalique /ae, i, u, ae:, i:, u:/ de l'ASM possède deux variantes combinatoires en fonction du contexte consonantique emphatique et non emphatique. Par ailleurs, du fait que les résultats de l'expérience perceptive de discrimination de timbre ne sont pas solidaires avec ceux de l'analyse acoustico-phonétique, nous pouvons affirmer qu'ils ne vont pas dans l'optique de la théorie motrice qui postule : il y a une relation étroite entre la production et la perception de la parole. Nos conclusions peuvent donc trouver une explication psychoacoustique en fonction du degré de rapprochement des deux premiers formants des voyelles de l'ASM. Cette explication correspond aux postulats de la théorie auditive qui considère que la perception de la parole ne se réfère pas aux connaissances articulatoires mais à des mécanismes du système auditif.

6. BIBLIOGRAPHIE

Amerman, J.D. & Daniloff, R.G. (1977) "Aspects of lingual coarticulation,"*Journal of Phonetics* 5, 107-13.

Badreddine, B. (1977) *Analyse phonologique et Phonétique du Parler de Kairouan.* Thèse de 3ème cycle. Paris X.

Belkadi, Y. (1984) Les voyelles de l'Arabe Standard Moderne: analyse spectrographique, *Travaux de l'Institut de Phonétique de Strasbourg*, 16, 217-240.

Chistovich, L. A. & Lublinskaya, V. (1979) "The center of gravity effect in vowel spectra and critical distance between formants: Psycho-acoustical study of the perception of vowel-like stimuli," *Hearing Research* 1, 185-195.

Ghazeli, S. (1977) *Back Consonants and Backing Articulation in Arabic.* Ph.D. Dissertation, University of Texax, Austin.

Kiel, N. (1987) A phonetic study of emphasis and vowels in Egyptian arabic, *Working Papers Lund University*, 30, 1-119.

Liberman, A. M & Mattingly, I. G. (1985) "The motor theory of speech perception revised," *Cognition*, 21, 1-36.

Metoui, (1989) *Contribution à la phonologie et la phonétique arabe : Etudes articulatoire et acoustique des voyelles du Parler de Tunis*, Idstein Schult-Kircher,Verlag.

Rajouani, A., Najim, M. and Chiadmi, D. (1987) "Synthesis of the pharyngealisation feature in arabic," *Speech Communication* 6, 261-268.

Schwartz, J.L. (1987) "A propos des notions de forme et de stabilité dans la perception des voyelles," *Bulettin du Laboratoire de la Communication Parlée*, Vol. 1A, 159-190.

Znagui, I. (1995) *Etudes phonétique et perceptive des voyelles de l'Arabe Standard Moderne.* Thèse de Doctorat, Université Paris III, Paris, 1-250.

LE ROLE DE LA SYLLABE DANS LA SEGMENTATION DES MOTS PARLES EN ITALIEN

Caroline FLOCCIA [1,3], Régine KOLINSKY[1, 2], José MORAIS[1]

1: Laboratoire de Psychologie Expérimentale, ULB, Bruxelles
2: Fonds National de la Recherche Scientifique, Belgique
3: Laboratoire de Psycholinguistique Expérimentale, FAPSE, Genève

ABSTRACT

Previous studies using fragment detection tasks in spoken words show that native listeners of syllable-timed languages such as French or Spanish rely on the syllable to segment the continuous speech signal. However, recent data obtained with Spanish suggest that the syllable segmentation process can be bypassed depending on the acoustic transparency of the language and the stimuli, a notion refering to the ease with which a segment can be identified as different from a competing candidate in a language. An experiment exploring fragment detection in Italian is presented. Italian like Spanish is expected to be a very acoustically transparent language, yielding similar results than those obtained in Spanish. However, results only partially replicate the Spanish data. The mandatory aspect of syllabic segmentation in syllable-timed languages is discussed.

INTRODUCTION

L'une des premières tâches nécessaires à la compréhension d'un énoncé linguistique oral consiste à segmenter le flux continu de parole en mots. Il a été proposé que la plus petite unité de segmentation ne soit pas de la taille du mot, mais de la taille de la syllabe (Mehler et al., 1981). Cependant, la nature de cette unité pré-lexicale - car elle interviendrait avant la reconnaissance des mots - semble varier selon la langue maternelle des sujets. En français (Mehler et al., 1981), comme en espagnol (Sebastiàn-Gallés et al., 1992; Bradley et al. 1993), en catalan (Sebastiàn-Gallés et al., 1992) ou en portugais (Morais et al., 1989), qui sont des langues décrites comme « syllable-timed », la syllabe semble effectivement jouer un rôle dans la segmentation des mots parlés. Ce n'est pas le cas en anglais (Cutler et al., 1986; Bradley et al., 1993), langue dans laquelle le rythme est basé sur une alternance entre syllabes fortes et faibles (« stress-timed »).

Ces conclusions sont basées sur des résultats obtenus grâce à des tâches de détection de séquences-cibles dans des mots de la langue: si la cible à détecter correspond exactement à la première syllabe du mot suivant (/ba/ dans /balance/ ou /bal/ dans /balcon/), on observe que les sujets français par exemple, sont plus rapides que lorsque la séquence ne correspond pas à la première syllabe du mot (/ba/ dans /balcon/ ou /bal/ dans /balance/) (Mehler et al., 1981). Cette interaction entre nature de la séquence et structure syllabique du mot est prise par ces auteurs comme preuve expérimentale que la segmentation d'un mot s'effectue syllabe par syllabe.

Cependant, l'effet syllabique, qui ne devrait dépendre a priori que de la structure rythmique de la langue, semble être fonction de plusieurs autres paramètres, dont la vitesse de réaction des sujets (Sebastiàn-Gallés et al., 1992), la position de l'accent dans le mot (id), le degré d'ambiguïté de la frontière syllabique entre première et deuxième syllabe des mots présentés (Zwitserlood et al., 1993). Sebastiàn-Gallés et al. (1992) ont proposé, pour les langues dites « syllable-timed », que la stratégie de segmentation syllabique pouvait être court-circuitée lorsque la *transparence acoustico-phonétique* de la langue ou des items présentés l'autorisait: dans ces cas-là, les sujets pourraient répondre très rapidement sur la base d'une représentation acoustico-phonétique de la première partie des mots. La notion de transparence fait référence à la facilité avec laquelle un segment peut être identifié et distingué d'un autre segment compétiteur de la langue (Sebastiàn-Gallés et al., p.19). Cette transparence serait d'autant plus importante que la variabilité autorisée dans la langue serait faible. Elle dépendrait notamment du nombre de voyelles utilisées, de la présence de réduction vocalique, d'ambisyllabicité et d'accent lexical.

Pour tester cette hypothèse, nous avons réalisé une expérience de détection de séquences dans une autre langue « syllable-timed », l'italien. La transparence de l'italien (au sens de la définition proposée par Sebastiàn-Gallés et al.) est très proche de celle de l'espagnol: les deux langues n'utilisent qu'un nombre restreint de voyelles (7 en italien contre 5 en espagnol), il n'y a pas de réduction vocalique, et très peu ou pas d'ambisyllabicité. De plus, comme en espagnol, il existe en italien un accent à

fonction lexicale et à position variable, mais portant en général sur l'avant-dernière syllabe des mots, qui accroitrait d'autant la transparence acoustique des syllabes concernées.

Le dessin général de notre expérience est une réplication des expériences présentées par Sebastiàn-Gallés et al. (1992) en espagnol. Dans cette étude, si les temps de réaction des sujets sont suffisamment lents (de l'ordre de 600 ms, Expérience 3), les effets syllabiques émergent, quelle que soit la position de l'accent dans les mots, c'est-à-dire quelle que soit la transparence acoustique des portions d'items à analyser. Par contre, lorsque les sujets répondent très vite (autour de 350 ms, Expérience 2), les effets syllabiques disparaissent, quel que soit le pattern accentuel des mots présentés, ce qui est interprèté par les auteurs comme l'indice que les sujets peuvent court-circuiter la stratégie syllabique si la transparence de la langue l'autorise, et si la tâche le requiert. Afin de répliquer les effets syllabiques obtenus par Sebastiàn-Gallés et al., nous avons testé la détection de séquences initiales dans des mots dont soit la première syllabe, soit la seconde syllabe, étaient accentuées. Afin de ralentir les sujets et ainsi favoriser l'émergence d'effets syllabiques, nous avons introduit des mots-pièges ainsi que des items contenant une consonne géminée en troisième phonème, par exemple /barro/ ou /barrito/. La présence de ces items devrait permettre de réduire d'autant la transparence acoustique des mots présentés. Si les sujets italiens se comportent comme leurs homologues espagnols, nous devrions répliquer les résultats obtenus par Sebastiàn-Gallés et al. dans leur Expérience 3, à savoir un effet syllabique quelle que soit la position de l'accent.

METHODE

Sujets

Vingt-trois étudiants italiens natifs de Bologne, âgés en moyenne de 25 ans et demi, ont été testés à l'Université de Psychologie de Bologne. Les données de huit étudiants supplémentaires n'ont pas été gardées, car soit plus de 10% de leurs temps de réaction excédaient 1500 ms, limite supérieure autorisée (3 sujets), soit leur taux d'ommissions ou d'erreurs excédait 5% des items (3 sujets), soit pour non compréhension des instructions (2 sujets).

Matériel

Douze triplets expérimentaux ont été construits. Dans chaque triplet, les mots partageaient les trois premiers phonèmes. L'un des mots était constitué d'une syllabe initiale de type CV, comme /balia/, le deuxième d'une syllabe de type CVC, comme /baldo/, et le troisième contenait une consonne géminée, comme /ballo/. La moitié des triplets étaient accentués sur la première syllabe, l'autre moitié sur la deuxième. Afin de respecter le pattern accentuel le plus fréquent en italien, les mots accentués en première syllabe étaient bisyllabiques, tandis que les mots accentués en deuxième syllabe étaient trisyllabiques. D'une catégorie accentuelle à l'autre, les triplets étaient appariés sur les trois premiers phonèmes. Par exemple les sujets devaient détecter /ba/ ou /bal/ dans les trois items /balia/, /baldo/ ou /ballo/, mais aussi dans /baleno/, /balcone/ et /balletto/. Les triplets commencaient par les séquences suivantes: /bar/, /bal/, /car/, /cal/, /mar/, /mal/, /mir/.

En plus de ces mots-test, dans lesquels il fallait détecter la séquence-cible associée, 72 items ont été choisis pour lesquels la cible associée ne correspondait pas à une séquence dans le mot. Pour la moitié de ces items, la cible et le début du mot ne partageaient aucun phonème, comme /ba/ - /neve/, ou /cal/ - /tentare/. Pour l'autre moitié, la cible partageait un, voire deux phonèmes, avec le début du mot associé, comme /ma/ - /mossa/, /car/ - /tabella/ ou /bar/ - /banchetto/. Deux blocs de 144 couples cible-mot ont été constitués. Chaque couple était répété deux fois, une fois dans chaque bloc, mais précédé et suivi de couples différents.

Procédure

Les cibles ont été enregistrées par un Italien, et les mots par une Italienne, afin de maximiser le contraste entre cibles et mots. Ces items ont été digitalisés sur un PC, et délivrés au travers d'un casque stéréo. Il était demandé aux sujets par écrit de répondre le plus vite possible, en appuyant sur un bouton placé devant eux (avec la main de leur choix), uniquement si la séquence sonore prononcée par l'homme correspondait au début du mot prononcé par la femme. Aprés quelques couples cible-mot d'entraînement, l'ordre de passation des blocs était contrebalancé sur les sujets. L'expérience durait en moyenne 20 minutes.

RESULTATS

Les temps de réaction inférieurs à 200 ms ou supérieurs à 1500 ms ont été éliminés.

Sur la figure 1 sont présentés les temps de réaction moyennés sur l'ensemble des sujets, obtenus sur les items accentués en première syllabe, en fonction de la correspondance entre la structure syllabique des cibles (CV ou CVC)

et celle des mots (première syllabe CV ou CVC). La figure 2 montre les résultats portant sur les mots accentués en deuxième syllabe.

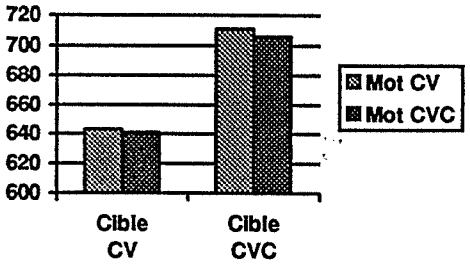


Figure 1: Temps de détection des cibles CV et CVC dans les mots accentués en première syllabe.

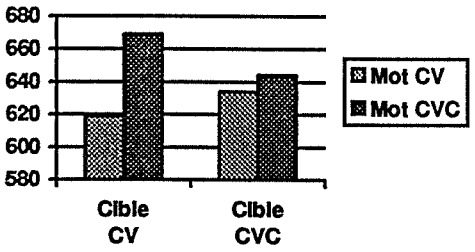


Figure 2: Temps de détection des cibles CV et CVC dans les mots accentués en deuxième syllabe.

Les temps de réaction moyens se situent autour de 665 ms. Globalement, les sujets sont plus rapides pour répondre sur les mots accentués en deuxième syllabe que sur les mots accentués en première syllabe (ANOVA sur la distribution des temps de réaction en fonction de la position syllabique, par sujet: $F(1,22)=4.66$; $p<.04$; par item: $F(1,11)=3.78$; $p=.08$).

Un effet syllabique semble émerger sur la figure 2, mais pas sur la figure 1.

Pour les mots accentués en première syllabe (figure 1), les sujets sont globalement plus rapides pour détecter les cibles de type CV que les cibles de type CVC (par sujet: $F(1,22)=23.6$, $p<.001$; par item: $F(1,11)=38.3$, $p<.001$). Les temps de réaction sont équivalents sur les mots de type CV ou CVC (par sujet, $F(1,22)<1$; par item: $F(1,11)<1$). Ainsi, lorsque la partie du mot sur laquelle doit se prendre la décision est acoustiquement transparente, on observe un avantage massif de la détection des cibles de type CV par rapport aux cibles de type CVC.

Pour les mots accentués en deuxième position (figure 2), l'interaction entre cibles et mots est significative par sujet ($F(1,22)=5.24$; $p=.03$) mais pas par item ($F(1,11)=1.73$). Le temps de détection des cibles de type CV est équivalent à celui des cibles CVC (par sujet: $F(1,22)<1$; par item: $F(1,11)<1$). Par contre, les temps de réaction sur les mots de type CV sont un peu plus rapides que sur les mots de type CVC (par sujet: $F(1,22)=3.89$; $p=.06$; par item: $F(1,11)=2.04$). Ainsi, lorsque la transparence acoustique de la partie du mot sur laquelle doit se prendre la décision est moindre, un effet syllabique semble se dessiner.

DISCUSSION ET CONCLUSION

Les résultats de cette expérience fournissent une indication supplémentaire que dans une langue habituellement décrite comme étant « syllable-timed », les sujets peuvent utiliser une stratégie de segmentation syllabique pour reconnaître des mots parlés.

Cependant, contrairement à ce qui a été observé par Sebastiàn-Gallés et al. (1992) en espagnol avec des temps de réaction aussi lents que les nôtres, nous n'obtenons pas d'effets syllabiques quelle que soit la position de l'accent dans les mots, mais seulement lorsque la première syllabe n'est pas accentuée. Ce pattern de résultats est par contre très proche de ce qui a été observé par ces mêmes auteurs avec le catalan, à savoir des effets syllabiques seulement lorsque la première syllabe des mots n'est pas accentuée, et une détection plus rapide des cibles CV que des cibles CVC dans les mots accentués en première syllabe.

L'italien serait-il plus proche du catalan que de l'espagnol, du point de vue de la transparence acoustique? Il est vrai qu'il existe 7 voyelles en italien contre 8 en catalan et 5 en espagnol. Mais sur d'autres points qui définissent la transparence acoustique d'une langue, notamment le peu de variabilité de la position de l'accent, et l'absence de réduction vocalique, l'italien reste beaucoup plus proche de l'espagnol que du catalan, et donc demeure une langue très transparente.

Il existe cependant une différence importante entre nos résultats et ceux de Sebastiàn-Gallés et al. en catalan, sur laquelle nous allons orienter la discussion: les temps de réaction moyens des sujets catalans sont presque deux fois plus rapides que ceux de nos sujets italiens.

Faisons l'hypothèse que d'une expérience de détection de séquence à l'autre, les temps de réaction soient des mesures comparables. Nous nous trouvons alors en face d'un résultat paradoxal par rapport aux hypothèses et aux résultats de Sebastiàn-Gallés et al.: rappelons que lorsque les temps de réaction de leurs sujets atteignent des valeurs comparables à celles que nous avons recueillies (supérieures à

600 ms), les effets syllabiques émergent, quelle que soit la position de l'accent dans le mot, en espagnol (Expérience 3 de Sebastiàn-Gallés et al.; voir aussi Bradley et al., 1993) comme en catalan (cité dans Kearns, 1994). D'après les auteurs, ces résultats sont l'indice que pour les langues dites « syllable-timed », le processus de segmentation syllabique est automatique et systématique, à partir du moment où *une certaine quantité d'information* a été traitée par le système perceptif.

Les temps de réaction très lents que nous avons recueillis correspondent à une réponse donnée lorsqu'une grande partie des mots a été traitée; notamment, pour les mots accentués en première syllabe, qui sont des mots bisyllabiques, les réponses des sujets ont même souvent été données *après* la fin des mots. Pourtant, dans ce cas précis, nous observons un processus de segmentation basé sur une représentation acoustico-phonétique du début des mots, ce qui signale que, même lorsqu'une *certaine quantité d'information* a été traitée, les sujets italiens peuvent court-circuiter la stratégie de segmentation syllabique pour donner leur réponse.

Si les Italiens peuvent répondre avant d'avoir appliqué une stratégie de segmentation syllabique lorsque la partie du mot sur laquelle doit se prendre la décision est accentuée, est-ce parce que la réalisation acoustique de l'accentuation en italien se caractérise par une plus grande transparence acoustique que dans les autres langues étudiées? Peut-être y-a-t-il plus de contraste entre une syllabe accentuée et une syllabe non accentuée en italien qu'en espagnol ou en catalan, notamment du point de vue de la longueur. Les locuteurs italiens disposant d'indices acoustiques plus saillants pour identifier les phonèmes, leur système perceptif pourrait donc s'appuyer sur une représentation sub-syllabique plus souvent que dans une autre langue. Des mesures sur le signal seront bien entendu nécessaires pour étayer ces arguments. Mais d'ores et déjà, des résultats obtenus avec des sujets français sur le matériel italien que nous avons décrit ici, permettent de renforcer notre hypothèse: quelle que soit la position de l'accent dans les mots italiens, les sujets français ne montrent aucun effet syllabique, mais ont des temps de détection plus rapides pour les cibles CV que pour les cibles CVC. Ainsi, la transparence acoustique des stimuli italiens est peut-être telle que, même pour des sujets aussi enclins à utiliser une stratégie syllabique que le sont a priori les francophones (voir Cutler et al., 1986), l'utilisation d'une représentation acoustico-phonétique est suffisante.

En résumé, nos résultats s'accordent avec une partie des conclusions de Sebastiàn-Gallés et al. (1992), notamment en ce qui concerne l'influence de la transparence d'une langue ou de certains stimuli, sur la possibilité pour les sujets de court-circuiter l'étape de segmentation syllabique dans une tâche de détection d'une cible dans le début d'un mot. Cependant, contrairement à ces auteurs, nous montrons qu'un tel processus peut avoir lieu à un moment tardif du traitement, ce qui pourrait remettre en cause l'aspect automatique et systématique de la segmentation syllabique des mots parlés. Pour valider cette hypothèse, il faudra explorer plus avant l'influence des valeurs des temps de réaction sur la direction des résultats dans le cas de l'italien. Notamment, il faudra déterminer si, lorsque les réponses des sujets sont encore ralenties, des effets syllabiques peuvent être mis en évidence, quelle que soit la position de l'accent dans les mots.

REMERCIEMENTS

Cette recherche a pu être mené grâce à un contrat Human Capital and Mobility ERB-CHRX-CT92-0031. Nous remercions Alessandra Sansavini et Elisabeta Ladavas pour leur assistance à Bologne, Uli Frauenfelder et Christine Meunier pour leurs commentaires et suggestions. Toute correspondance peut être adressée au premier Auteur, Laboratoire de Psycholinguistique, FPSE, 9, rte de Drize, 1227 Carouge, Suisse. Tel: (41 22) 705 97 41. E-mail: floccia@fapse.unige.ch.

BIBLIOGRAPHIE

Bradley D.C., Sànchez-Casas R.M. & Garcìa-Albea J.E. (1993) The status of the syllable in the perception of Spanish and English, *Language and Cognitive Processes*, n° 8, 197-233.

Cutler A., Mehler J., Norris D. & Segui J. (1986) The syllable's differing role in the segmentation of French and English, *Journal of Memory and Language*, n° 25, 385-400.

Kearns R. K. (1994) *Prelexical speech processing by mono- and bilinguals*, Thèse non publiée, University of Cambridge.

Mehler J., Dommergues J.Y., Frauenfelder U. & Segui J. (1981) The syllable's role in speech segmentation, *Journal of Verbal Learning and Verbal Behavior*, n° 20, 298-305.

Morais J., Content A., Cary L., Mehler J. & Segui J. (1989) Syllabic segmentation and literacy. *Language and Cognitive Processes*, n° 4, 57-67.

Sebastiàn-Gallés N., Dupoux E., Segui J. & Mehler J. (1992) Contrasting syllabic effects in Catalan and Spanish, *Journal of Memory and Language*, n° 31, 18-32.

Zwitserlood P., Schriefers H., Lahiri A. & van Donselaar W. (1993) The role of syllables in the perception of spoken Dutch, *Journal of Experimental Psychology: Learning, Memory and Cognition*, n° 19, 260-271.

LES EFFETS ASYMETRIQUES DE L'ACCENT LEXICAL SUR L'ACCES AU LEXIQUE CHEZ LE BILINGUE ARABE-ANGLAIS

Sami BOUDELAA

Université, PARIS 7
Laboratoire de Phonétique
10 Rue Charles V, 75004, PARIS.

Université Paris 5
Laboratoire de Psychologie Expérimentale
28 Rue Serpente, 75006, PARIS.

Abstract

Two lexical decision experiments examined the effects of lexical stress on word processing in Arabic-English bilinguals. In experiment 1, Arabic and English minimal stress pairs served as primes either to semantically related targets, to targets related to the second member of the pair, or to control targets. English minimal stress pairs were processed like homophones, but Arabic ones were not. In experiment 2, the effects of mis-stressing Strong-Weak and Weak-Strong common words (i.e. words that are not members of a minimal stress pair) was investigated. Only realizing a /SW/ word in a /WS/ stress pattern was adverse in English. In Arabic, however, mis-stressing had an adverse effect both in the case of SW and WS words. Taken together the results suggest (a) that the time course of lexical stress effects are *language dependent*, and (b) that bilinguals function monolingually with respect to lexical stress information. These results are explained in terms of the asymmetry underlying the phonological structure of the two languages.

Key words: stress, lexical access, bilinguals, phonological structure.

1. Introduction

Les recherches expérimentales sur le rôle de l'accent dans différentes langues offrent des résultats peu conciliables. Cutler & Clifton (1984) ont montré que la fausse accentuation des mots dissyllabiques ne retarde le traitement que lorsqu'un un mot originalement Strong-Weak ("SW") est réalisé avec un pattern accentuel Weak-Strong ("WS") (ex: minute réalisé *minute). Cutler (1986) a pu montrer que les paires minimales accentuelles -mots ne différant que par la place l'accent- sont traitées comme des homophones. En néerlandais, les résultats des expériences du "gating" montrent que l'opposition syllabe accentuée/syllabe non-accentuée est utilisée au cours des étapes précoces de traitement (Van Heuven, 1988). En revanche, les résultats des expériences d'amorçage intermodal dans cette même langue suggèrent que l'accent lexical serait utilisé d'une manière tardive (Jongenburger & Van Heuven, 1995). En arabe, des expériences d'amorçage intramodal (auditif-auditif) ont montré que l'accent peut être utilisé dès l'étape de l'activation : la fausse accentuation d'un mot retarde sa reconnaissance quel que soit le schéma accentuel du mot (Boudelaa, 1995).

Compte tenu de la disparité entre les données expérimentales relatives à l'effet de l'accent, la question de déterminer la nature exacte de l'information utile à la mise en relation entre le stimulus et la représentation interne reste posée : s'agit-il d'un appariement en termes de *phonèmes* (Marslen-Wilson, 1987, McClelland & Elman, 1986), en termes de *traits distinctifs* (Marslen-Wilson & Warren, 1994), ou d'un appariement en *parallèle* entre l'information *segmentale plus l'information prosodique* et une représentation interne (McQueen, Norris, Cutler, 1994, Banel & Bacri, 1994)?

Une façon d'éclaircir ce problème peut être envisagée dans le cadre de l'étude du bilingue, c'est-à-dire le sujet qui est capable de parler plus d'une langue en cas de besoin. L'étude du sujet bilingue revêt un intérêt certain en ce sens qu'elle permet (a) de comprendre les interactions entre deux répertoires linguistiques et (b) de déterminer la structure même du processeur linguistique en distinguant ce qui est *spécifique* à une langue donnée de ce qui a un *caractère universel* (Cutler, Mehler, Norris & Segui, 1992). Dans la présente recherche, on étudiera *le locus de l'effet de la syllabe accentuée lors la reconnaissance des mots isolés chez le bilingue arabe-anglais*. Il s'agit plus exactement d'examiner le rôle de l'accent lexical chez des bilingues arabes-anglais en utilisant le paradigme d'amorçage sémantique avec décision lexicale.

2. EXPERIENCE 1

Si l'information relative à la syllabe est utilisée au cours de l'étape de l'activation des candidats lexicaux chez le bilingue arabe-anglais, alors dans un paradigme d'amorçage sémantique une amorce "SW" membre d'une paire minimale (forbear) ne devrait faciliter que la cible qui lui est associée (ancestor). Le cas échéant, une amorce membre d'une paire

minimale faciliterait la cible qui lui est associée ainsi que celle associée à l'autre membre de la paire par rapport à une cible contrôle.

2.1. METHODE

Sujets : 16 bilingues arabe-anglais dominants en arabe, ainsi que 24 monolingues arabes et 24 monolingues anglais ont servi de sujets.

Matériel et procédure : Le matériel comprend 16 quadruplets arabes et 16 anglais. Le premier item de chaque quadruplet était un mot "SW" ou "WS" membre d'une paire minimale au sens de l'accent. Chaque membre de ces paires a servi d'amorce à :
- une cible qui lui est sémantiquement reliée (R1). ex : /wasʕafaa/-/ʃaraħa/, (ils ont décrit- il a expliqué), /wasʕafa/-/raaqa/ (il s'est éclaircit-il s'est amélioré) et 'forbear'-"ancestor", 'forbear'-"tolerate".
- une cible reliée à l'autre membre de la paire (R2).ex: /wasʕafa/-/raaqa/, /wasʕafa/-/ʃaraħa/ et 'forbear'-"tolerate", 'forbear'-"ancestor".
- une cible contrôle (C). ex: /wasʕafa/-/xaraʒa/ (il est sorti) /wasʕafa/-/naama/ (il s'est endormi) et 'forbear'-"arrival", 'forbear'-"vibrate".

Un test d'association lexicale préalable a permis de construire ces combinaisons. Ainsi, en arabe comme en anglais on a fait varier *l'Accent* ("SW"-"WS"), et la *Relation* entre amorce et cible (R1, R2 et C). De plus, 48 mots arabes et 48 mots anglais contrôlés pour la fréquence d'usage et la structure syllabique ont été sélectionnés. Ceux-ci ont servi d'amorces à des non-mots légaux. On a construit 6 listes de décision lexicale de 48 items contenant, chacune une cible R1, une cible R2 ou une cible C. Les mêmes amorces ou cibles n'apparaissent jamais plus d'une fois dans une même liste. Chaque bilingue devait accomplir une tâche de décision lexicale. Les deux groupes monolingues arabes et anglais ont entendu les stimuli de leur langue maternelle respective. Les temps de réaction étaient mesurés depuis la fin de la cible.

2.2. Résultats et discussion

Le tableau (I) présente les temps de réponses moyens (TR) et les (taux d'erreurs) des bilingues arabe-anglais et des monolingues arabes et anglais aux stimuli mots

Table 1: Temps de réponse moyens en ms et (écarts types) en fonction des différentes combinaisons du pattern accentuel avec le type de cible chez les bilingues avec les items arabes (Bil/ar) et les items anglais (Bil/ang), les monolingues arabes (M/ar) et les monolingues anglais (M/ang).

	Amorces SW			Amorces WS		
Sujets	R1	R2	C	R1	R2	C
Bil/ar	342 (37)	357 (44)	401 (42)	353 (43)	399 (38)	408 (38)
M/ar	342 (45)	355 (42)	393 (39)	350 (59)	390 (44)	395 (47)
Bil/ang	363 (35)	366 (44)	406 (43)	369 (45)	347 (43)	403 (40)
M/ang	352 (34)	359 (44)	398 (45)	350 (43)	363 (43)	396 (39)

L'analyse de variance des TR des bilingues pour le matériel anglais montre que l'effet de *l'accent* n'est pas significatif (F<1): une amorce facilite la cible qui lui est associée mais aussi celle associée à son partenaire accentuel. En revanche, le facteur *relation* avait un effet significatif (p<0,05), les cibles reliées "ancestor" et "tolerate" étant plus rapidement reconnus que les mots contrôles (p <0,05). L'interaction entre ces deux facteurs n'est pas significative (F<1). Le même résultat ressort de l'analyse des données des monolingues anglais. Ces résultats montrent que lors de la reconnaissance du mot en anglais chez le bilingue, l'accent lexical n'est utilisée que d'une manière tardive. En ce qui concerne l'arabe, l'analyse des TR des bilingues montre que l'effet de l'accent n'est pas significatif (F<1). Le facteur *relation* a un effet significatif (p < 0,05), les cibles reliées sont reconnues plus facilement que les cibles contrôles. Les deux facteurs interagissent significativement (p < 0,05) : une amorce "SW" comme /wasʕafa/ facilite la cible /ʃaraħa/ qui lui est associée ainsi que /raaqa/ qui est associé au "WS" /wasʕafa/; une amorce "WS" du type /wasʕafa/ ne facilite que la cible qui lui est associée. Les monolingues arabes ont révélé une performance similaire à celle des bilingues.

Un premier aspect important des résultats est que *le processeur linguistique ne tient compte que de l'information segmentale lors de l'appariement entre input perceptif et représentation interne*. Que l'amorce soit accentuée "SW" ou "WS" ne semble pas être pertinent au cours de l'activation des candidats pourvu que l'information segmentale ait une bonne correspondance avec une représentation lexicale (Marslen-Wilson, 1987). Il faut concéder toutefois, que le fait que les mots "WS" arabes membres d'une paire minimale ne facilitent que la cible appropriée indique que l'accent est *susceptible d'influencer le processus de la reconnaissance dès le début du*

traitement. L'asymétrie entre le traitement des paires minimales accentuelles arabes et anglaises est peut être due à l'absence d'équivalence entre les structures morphologiques des mots arabes qui forment des paires minimales. En effet, contrairement à l'anglais, un mot "SW" arabe correspond forcément à un préfixe plus une racine, par exemple /wa+sˤafa/.

Un deuxième aspect tout aussi important des résultats est que les bilingues arabes-anglais se comportent comme des monolingues arabes avec les données arabes et comme des monolingues anglais avec les données anglaises (Cutler et al., 1992., Grosjean, 1994). Ceci suggère que lorsque le bilingue traite un phénomène qui est asymétrique dans ses deux langues, il fait preuve d'un comportement *fonctionnellement monolingue*.

Afin de mieux élucider ces deux aspects des résultats -le rôle exact de l'accent lexical au cours du traitement et l'asymétrie et/ou la symétrie fonctionnel du bilingue- on va examiner l'effet de la fausse accentuation.

3. EXPERIENCE 2

Cette expérience étudie les effets de la fausse accentuation. Si l'accent lexical n'est utilisé que d'une façon tardive lors du traitement, alors le fait de réaliser l'accent d'une façon non orthodoxe dans des mots arabes et anglais qui ne sont pas membres des paires accentuelles, ne devrait pas gêner le traitement, à condition que la spécification segmentale du mot reste intacte.

3.1. METHODE

Sujets : Les mêmes que dans l'expérience 1.

Matériel et procédure : Les stimuli expérimentaux étaient constitués de 16 triplets de mots arabes et 16 triplets anglais. Le premier item de chaque triplet était un mot commun "SW" ou "SW" qui était réalisé avec un accent correct (AC), et un accent faux (AF). Un mot commun est considéré comme ayant un accent faux, lorsque ce dernier est attiré vers la droite si le pattern accentuel original est "SW" (ex: /dˤaraba/ "il a frappé" réalisé */dˤaraba/), ou vers la gauche si le pattern original est "SW". (/ʃaahadnaa/ "nous avons vu" est réalisé comme */ʃaahadnaa/. En anglais, le "SW" 'touchy' (susceptible) est réalisé comme "*touchy"; et le "SW" 'minute' (infime) est réalisé "*minute". Les versions AC et AF ont été utilisées pour amorcer des cibles reliées (R) et des cibles contrôles C choisies sur la base d'un test d'association préalable. De plus, 32 mots arabes et 32 mots anglais contrôles ont été sélectionnés. Ceux-ci ont servi d'amorces à des non-mots cibles légaux. La procédure de passation de l'expérience est identique à celle utilisée dans l'expérience 1.

3.2. Résultats et discussion

Les taux d'erreur ne dépassent jamais 5%. Les données anglaises montrent que chez le bilingue, une cible reliée (R) à un mot WS est reconnue plus rapidement (305 ms) qu'une cible C (375 ms) même si l'amorce est mal accentuée. Une cible R à un mot SW est reconnue plus rapidement (312 ms) que la cible C (398 ms) seulement si l'amorce est correctement accentuée (405 ms). En arabe, si la cible R est précédée de l'amorce correctement accentuée, elle est plus rapidement reconnue (SW 305 ms, WS 317 ms) que la cible contrôle (SW 405 ms, WS 398 ms). Si l'amorce est mal accentuée, alors les cibles R ont des TR aussi élevés que les mots C. L'analyse de variance des performances des bilingues avec les mots anglais montrent que ni le facteur "accent" ni le facteur "relation" n'avaient d'effet significatif ($F<1$). En revanche, ces deux facteurs interagissent significativement ($p<0,05$) : lorsque le bilingue arabe-anglais traite un input anglais mal accentué seul le déplacement de l'accent vers la fin du mot a un effet. Ces résultats sont les mêmes que ceux trouvés avec les monolingues anglais. Quant aux performances bilingues avec le matériel arabe, les effets du facteur "accent" ainsi que ceux du facteur "relation" étaient significatifs ($p < 0,05$). L'interaction entre ces deux facteurs n'était pas significative ($F< 1$): la fausse accentuation retarde le traitement des mots en arabe. Ces résultats sont analogues à ceux des monolingues arabes. Ainsi, en arabe *le moment* pendant lequel l'accent est prise en compte lors du traitement correspond à *l'activation des candidats lexicaux*. En revanche, en anglais cette information n'est considérée que lors de la *vérification* des candidats déjà sélectionnés sur la base de l'information segmentale.

4. Conclusion

Le bilingue arabe-anglais peut être *fonctionnellement monolingue* quant à l'utilisation de l'accent au cours de l'accès au lexique. Chez ce sujet, en arabe l'appariement entre *input auditif* et *représentation interne* se fait en tenant compte des informations segmentales et prosodiques. En revanche, en anglais le bilingue arabe-anglais ne tient compte que de l'information segmentale au cours des étapes précoces du traitement, gardant ainsi

l'information relative à l'accent lexical à une étape plus tardive. Cette asymétrie du rôle de l'accent s'explique (a) par la différence entre les informations acheminées par la syllabe accentuée dans les deux langues, et (b) par la différence entre les statuts de l'accent. En arabe, la syllabe accentuée fournit des renseignements sur la structure du mot ainsi que sur les structures des syllabes qui les composent. Ainsi, si on sait que la syllabe accentuée dans un mot est CV, on est sûr que ce mot là ne contient pas des structures syllabiques plus *lourdes* que CV. Aussi, la place de la syllabe accentuée est beaucoup moins imprévisible en arabe qu'en anglais. Or, la syllabe accentuée peut-être plus importante dans les langues à accent prévisible ou fixe que dans les langues où l'accent est complètement imprévisible (Pitt et Samuel,1990). A ce propos, les locuteurs du français exploitent activement l'opposition syllabe longue/syllabe brève (Banel et Bacri, 1994). En ce qui concerne l'anglais, la syllabe accentuée n'est importante que par opposition à une syllabe non-accentuée à voyelle réduite (McQueen et al., 1994). L'absence d'effet de l'accent dans la présente étude est probablement due au fait que les mots anglais utilisés n'ont pas de voyelle réduite. Rendre compte de l'ensemble de ces résultats nécessite que soit posé que le processeur linguistique humain accomplit la tâche d'appariement entre "input" et "représentation interne" en tenant compte de l'information segmentale et de l'information prosodique d'une façon plus ou moins parallèle en fonction de la langue traitée.

5. BIBLIOGRAPHIE

Banel, M.-H. & Bacri, N. (1994). On metrical patterns and lexical parsing in French: *Speech Communication, 15*, 115-126.

Boudelaa, S. (1995). The use of prosodic information in word recognition in modern standard Arabic: *Proceedings of ICPHS, Vol IV*, 340-343. Stockholm Sweden.

Cutler, A. & Clifton, C. (1984). The use of prosodic information in word recognition. in H, Bouma & D, Bouwhis (eds), Attention and performance: X. Control of language processes, 183-196. Hillsdale, NJ.

Cutler, A. (1986). Forbear is a homophone: Lexical prosody does not constrain lexical access, *Language & Speech, 3*: 201-219.

Cutler, A., Meheler, J. Norris, D. & Segui, J. (1992). The monolingual nature of speech segmentation by bilinguals, *Cognitive Psychology, 24*, 381-410.

Grosjean, F. (1994). Going in and out of languages: An example of bilingual flexibility: *Psychological Sciences 4*, 201-206.

Jongenburger, W. & Van Heuven, V. (1995). The role of linguistic stress in the time course of word recognition in stress-accent languages: *Proceedings of Eurospeech, Vol III*, 1695-1698. Spain.

Marslen-Wilson, W. D. & Warren, P. (1994). Levels of perceptual representation and process in lexical access: *Psychological Review, 4*, 653-675.

Marslen-Wilson, W. D. (1987). Functional parallelism in spoken word recognition. (in) U. H. Frauenfelder & L. K. Tyler (eds) *Spoken word recognition. A cognition special issue*. MIT. Press.

McClelland, J. L. & Elman. J. L. (1986). The TRACE model of speech perception *Cognitive Psychology, 18*, 1-86.

McQueen, J., Norris, D. & Cutler, A. (1994). Competition in spoken word recognition: Spotting words in other words, *J. Experimental Psychology: Learning, Memory and Cognition, 20*, 621-638.

Pitt, M. A. & Samuel, A. G. (1990). The use of rythem in attending to speech: *J. Expériemental Psychology: Human Perception and Performance, 3*, 564-573.

Van Heuven, V. (1988). Effects of stress and accent on the human recognition of word fragments in spoken context: Gating & shadowing, *FASE Symposium*, 811-818, Edingburgh.

RECONNAISSANCE PAR DES LOCUTEURS MONOLINGUES ET BILINGUES (ESPAGNOLS ET FRANÇAIS) DE L'ESPAGNOL ET DU FRANÇAIS À PARTIR DE PRODUCTIONS FILTRÉES.

Martine Le Besnerais

Universitat Autònoma de Barcelona - Fac. de Lletres – Dep. de Fil. Francesa i Romànica - 08193 Bellaterra (Spain)

Tel.: (343) 581 14 10 - Fax: (343) 581 20 01 - e-mail: martine@prosodia.uab.es

ABSTRACT

The processes of speech perception and speech production interact under the constraint of communicative goals. This interaction is as caracteristic of prosodic processing as of the processing of other aspects of linguistic structure. In this paper, we tried to evaluate the incidence of prosodic factors (essentially rhythmic factors) on the level of language identification. The high scores of identification demonstrated prosodic components play a great role in the task of identification of languages. Bilingual subjects got higher scores than monolinguals. Nevertheless, all subjects, bilinguals and monolinguals, regardless of first or second language, got similar scores for both languages. Several issues that arise in such work are discussed, and suggestions are made for a different cognitive treatment of both categories of subjects.

1. INTRODUCTION

L'objectif des analyses perceptives dans le cadre des études sur la parole peut différer selon les approches, soit que l'on entende valider l'analyse expérimentale des données acoustiques par un test de perception qui corrobore, complète ou infirme les résultats, soit que l'on attende de l'interprétation perceptive du signal, l'information nécessaire à la catégorisation, soit enfin que l'on se donne comme objectif l'étude des mécanismes et processus mis en œuvre par le traitement perceptif. À ce sujet, les recherches les plus récentes ne considèrent pas la perception en tant que fonction directe entre un stimulus et sa forme perçue, mais comme le résultat d'un processus de traitement dans lequel l'accent est mis sur les conditions de fonctionnement et l'effet des contextes, des situations et de l'expérience. Les processus perceptifs constituent une échelle continue de traitements depuis les automatismes ou spécialisation du traitement de l'information, jusqu'aux processus centraux qui opèrent de façon intégrée. Ces deux conceptions du traitement perceptif (en tant qu'activité essentiellement sensorielle ou en tant qu'activité cognitive de traitement), concernent directement l'analyse perceptive d'énoncés filtrés visant à l'identification de systèmes de langues à partir de la seule composante suprasegmentale qui sera présentée dans ce travail; on devra s'y référer pour l'interprétation des résultats.

2. OBJECTIFS

Les études (Allen, 1975) ont démontré le caractère universel des phénomènes de rythme non seulement parce qu'ils sont contraints à des lois de caractère biologique et psychologique mais également parce quils sont porteurs d'une fonction linguistique, principalement dans l'organisation de la syntaxe, et dans celle du message (division thème/rhème) opérant ainsi une hiérarchisation de l'information (Rossi, 1985). Paradoxalement, la caractérisation de la prosodie d'une langue en terme de dichotomie est essentiellement une caractérisation de son rythme; il s'ensuit que nous sommes en présence d'un phénomène à la fois universel et foncièrement discriminant. Notre propos a été celui de savoir si au niveau de la perception, la reconnaissance d'une langue pouvait être effectuée sur la seule base des éléments suprasegmentaux. La typologie prosodique des langues, qui porte essentiellement sur des dichotomies rythmiques, octroie à l'une et l'autre des deux langues, l'espagnol et le français, certaines caractéristiques essentielles communes : langues à rythmicité syllabique, "trailer-languages" (Wenk & Wioland, 1982), langues à structure syllabique CV : 58% CV pour l'espagnol et 56% CV pour le français (Dauer, 1982) et certaines caractéristiques distinctes: langue à accent libre discriminant

(espagnol) et langue à accent fixe démarcatif (français).

3. PERCEPTION : INTERFÉRENCES

Comprendre, interpréter les processus de perception du langage parlé suppose de bonnes connaissances des niveaux de fonctionnement hiérarchiques permettant à l'auditeur d'associer un message phonétique discret à un signal acoustique essentiellement continu.

Les mécanismes perceptifs qui interviennent dans notre faculté de reconnaître des variations dans le continuum sonore ont été étudiés au niveau segmental dans le cadre de recherches phonologiques intralangues, dans des approches contrastives interlangues, dans le cadre de recherches sur l'évolution de la faculté de conceptualisation au cours des différents stades de développement de l'individu, dans le cadre de l'acquisition d'une langue étrangère et des processus inférentiels, dans le cadre des recherches en synthèse et en reconnaissance automatique. On sait combien la structure phonologique de la langue des sujets s'impose dans les tâches de discrimination. On sait également qu'on ne peut appliquer sans réserve les modèles de discrimination **intralangue** à des modèles de discrimination **interlangue.** En outre, bien que l'on accorde de plus en plus de crédit aux hypothèses qui postulent une approche globale des phénomènes de décodage, le rôle des indices prosodiques dans les premières étapes de traitement du signal ont été étudiés dans une moindre mesure, exception faite des recherches en genèse du langage (Konopczynsky, 1986). L'étude précise de la perception des traits prosodiques est souvent faussée par la relation étroite qu'elles entretiennent avec d'autres informations linguistiques (syntaxiques et sémantiques). En ce sens, notre dispositif expérimental transgresse ces corrélations, mais l'interprétation des résultats reste difficile. En revanche, on pourra être à même de déterminer l'impact des structures suprasegmentales dans l'activation du système de recherche de l'information à partir des taux de reconnaissance que nous avons obtenus.

4. PROCÉDURE EXPÉRIMENTALE

Nous avons enregistré six productions en espagnol et six en français: deux locuteurs et une locutrice en espagnol et un locuteur et deux locutrices en français; la durée de chacun des enregistrements est d'environ 5 mn. qui comprennent une partie de texte lus et une partie de parole spontanée ou commentaire libre sur les thèmes abordés. Les énoncés ont été filtrés à l'aide de l'appareil SUVAG (la pente ou importance de l'atténuation est de 60 dB par octave) dont on a utilisé le filtre passe-bas à une fréquence de coupure atténuant les fréquences supérieures à 300 Hz. Cette fréquence de coupure comportait le risque d'une identification du français et de l'espagnol à partir du premier formant des phonèmes /i/, /u/ et surtout /y/ (qui n'appartient pas au système vocalique de l'espagnol) et qui se situe en français à 250 Hz et en espagnol à 275 Hz (Delattre, 1965); nous avons rejeté malgré tout l'autre option possible, c'est-à-dire de descendre à 200 Hz, car on a considéré l'importance de ne pas amputer les courbes de Fo de ses valeurs maximales, telles que nous les avions détectées sur les tracés. On a estimé que, puisqu'il existe une corrélation entre la fréquence fondamentale intrinsèque de la voyelle et la fréquence de son premier formant, on pouvait éviter cet écueil en choisissant pour le français au moins deux voix de femme ce qui supposait une augmentation de 18 à 20% pour les valeurs formantiques qui ne risquaient plus d'apparaître dans les productions filtrées (on sait par ailleurs que la reconnaissance d'une voyelle à partir de sa valeur intrinsèque et de son premier formant est tout à fait improbable). L'enregistrement de ces productions filtrées a été effectué sur la platine TASCAM 112 dont la fréquence de réponse se situe entre 25 Hz et 19 KHz (+/- 3 dB).

La reconnaissance de ces productions a été soumise à :
— 20 sujets monolingues français;
— 20 sujets monolingues espagnols;
— 20 sujets bilingues français-espagnol;
— 20 sujets bilingues espagnol-français;

Du signal original, il ne leur parvenait que la fréquence fondamentale, l'amplitude et les caractéristiques de l'organisation temporelle. Aucun des sujets monolingues ne méconnaissait totalement l'autre langue soumise à l'écoute. Tous les locuteurs avaient la possibilité de réaliser deux types de tâches perceptives : l'identification (espagnol *vs* français) et la mise à l'écart, la fiche remise comportant une colonne "ambigu". Enfin ils ignoraient le nombre de productions dans chacune des langues.

5. RÉSULTATS

Le taux de reconnaissance (tous sujets confondus) des productions en espagnol et en français est de 77%, ce qui constitue un résultat significatif du point de vue de la conscience que les sujets ont du rythme des langues soumises à examen. Les scores obtenus en fonction de la catégorie des locuteurs, en fonction de la langue à identifier et en fonction des types d'erreurs effectuées apparaissent dans le tableau suivant, sous forme de pourcentages.

Table 1: Résultats des tests de perception en fonction des quatre catégories de sujets et des réponses données à l'écoute des productions filtrées de l'espagnol et du français.

	Identif.Esp	Identif.Fr.	Er.sur Esp	Er. sur Fr.	Amb sur Esp.	Amb sur Fr.	Taux Identif.
Monolingues français	41 / 60 68%	45 / 60 75%	6 / 60 10%	9 / 60 15%	13 / 60 22%	6 / 60 10%	86 / 120 72%
Monolingues espagnols	39 / 60 65%	37 / 60 62%	14 / 60 23%	15 / 60 25%	7 / 60 12 %	8 / 60 13 %	76 / 120 63%
Bilingues français / espagnol	52 / 60 87%	50 / 60 83%	4 / 60 7%	8 / 60 13%	4 / 60 7%	2 / 60 3%	102 /120 85%
Bilingues espagnol / français	54 / 60 90%	51/ 60 85%	4 /60 7%	5 / 60 8%	2 / 60 3%	4 / 60 7%	105 / 120 88%
Taux Identification	186/ 240 77%	183/ 240 76%	28/240 12%	37 / 240 15%	26 / 240 11%	22 / 240 8%	369 / 480 **77%**

5.1. Résultats en fonction des deux catégories de sujets

Les résultats détaillés indiquent en ce qui concerne les auditeurs, de meilleurs scores pour les sujets bilingues (86%) que pour les sujets monolingues (68%) –ce qui confirme les résultats de (Ohala & Gilbert, 1979) sur la compétence des locuteurs monolingues, bilingues ou même trilingues– et également la tendance dans le cas des bilingues à écarter l'option "ambigu". Les scores obtenus sont, même dans le cas des sujets monolingues, des scores élevés d'identification. Il semble bien qu'il y ait donc des critères de sélection qui se fondent sur des différences physiques et/ou interprétables. Pourtant, les résultats, comme on le verra dans le paragraphe suivant, ne sont pas fonction du statut de la langue des auditeurs.

5.2. Analyse des résultats en fonction du statut de la langue

On observe une proportion de pourcentages de reconnaissance analogues, en ce qui concerne le statut de la langue des auditeurs : langue maternelle et "langue seconde", pour les deux catégories de sujets qui ont réalisé le test. Ceci diffère des résultats obtenus sur des corpus lus (Maidment, 1976) où les productions en langue 1 donnent lieu à des taux de reconnaissance plus élevés. Nos résultats ne confirment pas cette différence de scores. Il est clair que les sujets bilingues identifient sans difficultés les deux systèmes de langues. Cela est révélé par les pourcentages de réponses correctes dans l'une et l'autre des deux langues et également par les cas très peu nombreux de réponses "ambigu". Les sujets monolingues obtiennent des résultats inférieurs, également non différenciés sur le plan du statut de la langue (L1 ou L2); la confrontation s'établirait sur la discrimiation d'éléments connus et d'éléments inconnus. Les pourcentages de réponses (langue 1 et langue "autre") n'ont en principe pas lieu d'être déséquilibrés. En revanche, la tâche est plus difficile du fait que les éléments connus ne le sont qu'à travers une confrontation interne au système de langue.

5.3. Analyse des erreurs

En ce qui concerne les erreurs, on observe, qu'indépendamment de l'option "ambigu", elles sont plus nombreuses sur du français que sur de l' espagnol, et ce pour les quatre catégories de sujets. Elles sont, malgré tout, davantage le fait des monolingues espagnols que des monolingues français. Les taux les plus élevés de réponses "ambigus" sont donnés par les locuteurs monolingues français. Mais les taux d'erreurs sont insuffisants pour faire l'objet d'une interprétation détaillée. On peut cependant faire quelques remarques à ce propos :

L'auditeur monolingue possède une compétence langagière qui lui permet 1° de discriminer dans sa propre langue les marques phoniques pertinentes (par exemple /s/ et /z/ du français), 2° de ne pas discriminer unité phonologique et variantes sonores (par

exemple [s] et [z] de l'espagnol). En contact avec une langue étrangère les erreurs les plus fréquentes qu'engendre le système des cribles ont lieu au niveau de la perception d'une unité phonologique appartenant au système de la langue maternelle pour une autre appartenant au système de la langue étrangère. Mais il est probable qu'au niveau prosodique, des phénomènes de discrimination phonologique comme ceux évoqués plus haut (cf. 2°) aient une plus grande incidence sur la perception de l'organisation prosodique de L2. Cela signifierait que si certains indices de L2 ne sont pas pris en compte par les auditeurs monolingues c'est qu'ils ne sont pas pertinents en L1 (comme dans l'exemple de [s] et [z] de l'espagnol) et inversement, on peut considérer que si certains indices de L2 sont phonologisés à mauvais escient c'est qu'ils sont distinctifs en L1 (comme dans l'exemple de /s/ et /z/ du français).

En effet, selon R. Renard (1977), chacun des sujets parlants –quelle que soit sa langue– a sans doute expérimenté en production ou en perception la même gamme de schémas prosodiques, ce qui n'est pas le cas pour les traits phonémiques. Ceci apporterait une explication aux constatations qui ont émergé des études faites sur des productions filtrées et qui montrent que les taux de reconnaissance sont directement corrélés à la longueur des corpus présentés. Sans doute l'identification est-elle davantage liée à la fréquence d'emplois de schémas prosodiques et à leur valeur fonctionnelle qu'à la détection d'un archétype. Les résultats ne nous renseignent pas sur la nature de ces éléments qui ont occasionné les choix des sujets.

6. CONCLUSION

Les scores obtenus sont suffisamment élevés pour nous permettre de conclure que les éléments prosodiques, c'est-à-dire l'organisation sur l'axe temporel des variations de fréquence fondamentale, d'intensité et également des pauses constituent des « détecteurs » qui jouent un rôle essentiel dans la reconnaissance d'un système phonologique donné et ce, indépendamment de toute composante verbale. Les résultats le montrent mais leur interprétation en termes de traits ne peut être envisagée dans le cadre de ce travail. Dans le cas des sujets monolingues, (dont les scores sont inférieurs à ceux des sujets bilingues), on peut supposer que la mise en œuvre du crible phonologique engendre des interférences au niveau des unités fonctionnelles; mais on peut supposer également que la compétence prosodique des sujets neutralisera une opposition pertinente dans une langue parce que non pertinente dans la leur ou inversement "phonologisera" une réalisation acoustique tributaire du contexte parce que cette reálisation est pertinente dans leur système de langue. Les meilleurs résultats des sujets bilingues tendraient à prouver que, chez ces derniers, ces phénomènes d'interférences ne se présentent plus au niveau de la perception, ce qui confirme l'assertion de Grosjean (Grosjean, 1989) : *« Neurolinguists, Beware! The bilingual is not two monolinguals in one person ».*

7. BIBLIOGRAPHIE

Allen, G. D. (1975) Speech rhythm : Its relation to performance universals and articulatory timing, *J. of Phon.*, n°3, 75-86.

Cutler, A. (1994) The perception of rythm of language, *Cognition,* n°50, 79-81.

Dauer, R. M. (1982) Stress-timing and syllable-timing reanalyzed, *J. of Phon.*, n°11, 51-62.

Diana, A. (1994) Test de hauteur et perception de l'accent en anglais par des francophones : une première approche, *Actes des XXèmes Journées d'Étude sur la parole,* 401-406.

Delattre, P. (1965) *Comparing the Phonetic Features of English, French, German and Spanish,* An Interim Report, Heidelberg, Chilton, Groos.

Grosjean, F., (1989) Neurolinguists, Beware! The Bilingual is Not Two Monolinguals in one Person, *Brain and Language,* n° 36, 3-15.

Konopczynsky, G. (1986) *Du prélangage au langage : acquisition de la structuration prosodique,* Thèse de doctorat, Univ. Strasbourg.

Lea, W. A. (1974) Prosodic aids to speech recognition: a summary of results to date, *Sperry Univac Technical Report,* No PX 11087.

Maidment, J. (1976) Voice fundamental frequency characteristics as language differentiators, *Speech and Hearing Work in Progress,* University College, London 2, 74-93.

Ohala, J. J., Gilbert, J. B. (1979) Listeners´ ability to identify languages by their prosody, in Léon, P. & Rossi, M. (Dir) *Problèmes de Prosodie,* II, Coll. *Studia Phonetica,* n°18, Ottawa. Didier. 123-132.

Pasdeloup, V., Kolinsky, R. & Moras, J. (1994) La chaîne phonémique et le patron accentuel lexical sont-ils représenté séparément ? une étude d'illusions perceptives, *Actes des XXèmes Journées d'Etude sur la Parole,* Trégastel, 523-528.

Renard, R. (1977) Prosodie, phonémique : des stratégies différentes d'éducation de l'acte audio-phonatoire ?, *non publié.*

Rossi, M. (1985) L'intonation et l'organisation de l'énoncé, *Phonetica,* n°42, 135-153.

Wenk B.J., Wioland, F.(1982), Is French really syllable-timed?, *J. of Phon* n°10, 193-216

RÔLE DE LA STRUCTURE MORPHOLOGIQUE DANS LE TRAITEMENT DU LANGAGE PARLÉ

Fanny MEUNIER, Juan SEGUI

Université René Descartes, Laboratoire de Psychologie Expérimentale, URA 316 CNRS,
28 rue Serpente, 75006, Paris, France.
Tél.: 40 51 98 65 - Fax: 40 51 70 85 - fmeunier@ext.jussieu.fr

ABSTRACT

The research described here is concerned with lexical entry. It addresses the issue of how this is organised: whether lexical representations are word-based or morpheme-based. We studied two types of derived words separatly: prefixed and suffixed words. In this experiment we used a cross-modal priming paradigm. Different patterns of results were observed for prefixed and suffixed words. We make the assumption that this difference is due to the structural characteristics of the French language. We concluded that the lexical entries of prefixed word are word-based, whereas those of regular suffixed words are morpheme based.

Key words: morphological structure, lexical access representation, cross modal priming.

1. INTRODUCTION

Le but de ce travail est de fournir des informations sur le mode de traitement et de représentation des mots morphologiquement complexes.

Un nombre très important d'études empiriques et de cadres théoriques ont été développés sur ce sujet depuis une vingtaine d'années en Psycholinguistique (voir Mc Quenn & Cutler, 1994; Pillon, 1993).

D'après certains modèles théoriques, les relations morphologiques entre les mots sont encodées dans le lexique mental d'une manière spécifique non réductible aux relations formelles (orthographiques et/ou phonologiques) ou sémantiques entre ces mots (Taft & Forster, 1975). Autrement dit, pour ces théories, les relations morphologiques constituent un principe d'organisation fondamental du lexique interne (ex, Beauvillain et Segui, 1992).

La plupart des travaux destinés à la mise à l'épreuve de ce type d'hypothèse a été conduite à l'aide de la procédure de l'amorçage ou "priming". En conformité avec l'hypothèse morphologique, ces recherches ont montré, par exemple, que la présentation préalable d'un mot morphologiquement complexe (ex, "reprise") facilite le traitement ultérieur de sa composante racine (ex, "prise"). Cet effet a été attribué à la mise en oeuvre d'une procédure de décomposition du mot dérivé dans ces constituants morphémiques affixe et racine (Taft et Forster, 1975).

Plus généralement, on observe un effet de facilitation quand le mot-amorce et le mot-test sont issus d'une même famille morphologique. Ces résultats suggèrent que la présentation d'un mot appartenant à une famille morphologique "active" dans le lexique mental du sujet les autres membres de cette famille. Toutefois, l'importance des effets morphologiques de facilitation varie en fonction d'un grand nombre de paramètres dont les suivants : a) nature flexionnelle ou dérivationnelle de la relation, b) organisation interne des mots affixés (mots préfixés ou mots suffixés) (Colé, Beauvillain & Segui, 1989), c) modalité sensorielle de présentation des mots (modalité visuelle, modalité auditive, présentation intermodale auditive-visuelle) (Connine, Mullenix, Shernoff & Yelen, 1990).

Dans l'expérience présentée, nous avons abordé la problématique du traitement des mots morphologiquement complexes parlés en utilisant une procédure d'amorçage intermodale auditive-visuelle. Dans cette procédure le mot-amorce est présenté par voie auditive suivi immédiatement par la présentation visuelle du mot-cible. Cette procédure est particulièrement adaptée pour l'étude du traitement des mots présentés par voie auditive. Contrairement à la présentation visuelle, une présentation auditive exige un traitement séquentiel des constituants du mot.

Les relations morphologiques étudiées sont exclusivement des relations de dérivation; relation entre un mot dérivé et sa racine ou bien

entre deux mots dérivés issus d'une même famille.

Les principales questions soulevées dans cette recherche concernent les points suivants:

1) Rôle de la transparence phonologique du lien existant entre un mot-dérivé et sa racine. Il s'agit de savoir en particulier, si l'effet de facilitation est comparable pour des couples de type "brutal"-"brute" (relation transparente) et "chaleur"-"chaud" (relation opaque).

2) Rôle de la nature de la relation entre deux mots dérivés d'une même famille. Nous comparerons dans ce cas l'effet d'amorçage pour des couples de mots préfixés et pour des couples de mots suffixés. Cette comparaison doit nous permettre d'établir dans quelle mesure l'ordre des composants racine-affixe de ces deux sortes de mots peut affecter leur traitement. Par exemple: "refaire"-"défaire" et "jardinier"-"jardinage".

3) Rôle de la relation morphologique par rapport à la relation phonologique entre les mots de la paire. Il s'agit d'établir ici si l'effet de facilitation morphologique peut être réductible à un effet de relation phonologique. Pour répondre à cette question nous comparerons l'effet observé pour des couples reliés morphologiquement tels que "amoral" - "moral" avec des couples analogues du point de vue de la relation phonologique mais non reliés morphologiquement tels que "alarme"-"larme".

4) Rôle de la relation morphologique par rapport à une relation purement sémantique entre les mots. Nous chercherons à établir dans quelle mesure les effets morphologiques peuvent être réductibles à des effets sémantiques en comparant les effets observés pour des couples de mots reliés morphologiquement et sémantiquement. Par exemple: "brutal"-"brute" et "personne"-"individu".

2. EXPÉRIENCE

Dans notre expérience nous avons utilisé le paradigme d'amorçage inter-modal: les sujets entendent un mot amorce et à la fin de sa réalisation acoustique apparaît sur un écran un mot cible. Les sujets doivent effectuer une tâche de décision lexicale sur ce mot. On s'intéresse ici à la différence de temps d'identification du mot cible lorsque celui ci est précédé par une amorce liée et lorsqu'il est précédé par une amorce contrôle non liée.

2.1. Sujets

Trente-huit étudiants de Psychologie (ayant entre 18 et 30 ans) ont participé à l'expérience. Tous étaient de langue maternelle française, ne présentaient aucun trouble de l'audition et/ou de la vision.

2.2. Matériel expérimental

Nous avons utilisé des paires de mots expérimentaux constituées d'un mot dérivé et de sa racine, ou de deux mots dérivés d'une même racine. Comme nous l'avons déjà vu nous nous sommes intéressés aux mots préfixés et aux mot suffixés séparément. Des exemples d'amorces et de cibles sont présentés dans les tables 1, 2 et 3.

Table 1: Exemples de Mots Préfixés (M= liaison morphologique; S= liaison sémantique; P= liaison phonologique; R= racine; D= mot dérivé; Contrôle= amorce contrôle; Liée= amorce liée).

			Amorce	Cible
M+S+P+	D/R	Liée	impartial	partial
		Contrôle	colossale	
M+S+P-	D/R	Liée	imberbe	barbe
		Contrôle	manière	
M+S+P+	D/D	Liée	retourner	contourner
		Contrôle	passionner	
M-S-P+	D/R	Liée	adorer	dorer
		Contrôle	opérer	

Table 2: Exemples de Mots Suffixés (les abréviations et symboles sont identiques à ceux utilisés Table 1).

			Amorce	Cible
M+S+P+	D/R	Liée	brutal	brute
		Contrôle	subtil	
M+S+P-	D/R	Liée	circulaire	cercle
		Contrôle	majestueux	
M+S+P+	D/D	Liée	balayage	balayeur
		Contrôle	crustacé	
M-S-P+	D/R	Liée	louper	loup
		Contrôle	farcir	

Table 3: Exemple de synonimes (M= liaison morphologique; S= liaison sémantique; P= liaison phonologique; Syn= synonime).

			Amorce	Cible
M-S+P-	Syn	Liée	personne	individu
		Contrôle	autorité	

La première condition (M+P+S+) est constituée de couples de mots dérivés-racine phonologiquement transparents: la réalisation de la racine est la même, qu'elle soit libre ou associée à un affixe.

Dans une deuxième condition (M+P-S+), les relations phonologiques et orthographiques de la racine et de son dérivé sont opaques.

Une troisième condition (M+P+S+), correspond aux couples de mots dérivés d'une même racine.

Une quatrième condition (M-P+S-) est constituée de paires de mots mophologiquement non reliés, mais partageant les même relations phonologiques qu'une racine et un mot dérivé, le plus petit des mots est toujours une sous-chaîne du mot plus long.

La cinquième condition (M-P-S+) est constituée de paires de mots synonymes qui sont reliés sémantiquement.

Le matériel était constitué de 15 couples pour chacune des cinq conditions décrites.

Deux listes expérimentales ont été construites, de telle manière que chaque sujet voit la moitié des cibles de chaque condition dans une situation où l'amorce est liée à la cible et l'autre moitié dans une situation contrôle où l'amorce n'est pas liée à la racine (cible).

2.3. Procédure

Le sujet entend toujours le mot dérivé, puis il voit la cible et doit faire sa décision lexicale sur le mot écrit.

2.4. Résultats

Deux analyses de variance, par sujets et par items, ont été conduites séparément pour les deux types de mots (préfixés et suffixés).

Table 4 : Temps de décision lexicale moyennés en ms pour les Mots Préfixés (les abréviations et symboles sont identiques à ceux utilisés Table 1; * = différence significative à p<.05).

		Liée	Non Liée
M+S+P+	D/R Transp	538	568 *
M+S+P-	D/ROpaque	514	532 *
M+S+P+	D/D	608	643 *
M-S-P+	Phonolo	592	598

Table 5 : Temps de décision lexicale moyennés en ms pour les Mots Suffixés (les abréviations et symboles sont identiques à ceux utilisés Table 1; * = différence significative à p<.05).

		Liée	Non Liée
M+S+P+	D/R Transp	514	557 *
M+S+P-	D/ROpaque	517	521
M+S+P+	D/D	591	624 *
M-S-P+	Phonolo	588	591

Table 6: Temps de décision lexicale moyennés en ms pour les Synonimes (les abréviations et symboles sont identiques à ceux utilisés Table 3; * = différence significative à p<.05).

		Liée	Non Liée
M-S+P-	Synonymes	583	606*

Mots préfixés:

On trouve un effet d'amorçage entre un mot dérivé et sa racine que leurs relations phonologiques soient transparentes (30 ms) ou opaques (18 ms). Ces deux effets sont significatifs et par sujets et par items.

De plus on ne trouve pas d'effet d'amorçage entre deux mots non reliés morphologiquement mais qui partagent des relations phonologiques analogues (condition phonologique).

Enfin on trouve un effet d'amorçage significatif par sujets et par item entre deux mots préfixés dérivés d'une même racine (35 ms).

Mots suffixés:

On observe qu'un mot suffixé amorce sa racine (43 ms) mais uniquement si celle-ci est phonologiquement identique lorsqu'elle est libre et lorsqu'elle est associée à l'affixe.

Par ailleurs comme pour les mots suffixés, on n'observe pas d'amorçage entre deux mots ne partageant que des relations phonologiques.

De plus, les résultats traduisent un amorçage entre deux mots suffixés dérivés d'une même racine (33 ms).

Pour finir il est à noter qu'on trouve un amorçage sémantique (23 ms) entre deux synonymes.

3. DISCUSSION ET CONCLUSION

1- Pour les mots préfixés on observe donc un amorçage morphologique entre un mot dérivé et sa racine. Cet effet d'amorçage existe même lorsqu'il n'y a pas un recouvrement phonologique parfait entre la racine lorsqu'elle est libre et lorsqu'elle est intégrée à un dérivé.

Pour les mots suffixés, le pattern est différent: on observe un amorçage entre un mot suffixé et sa racine uniquement si leur relation morphologique est phonologiquement transparente. L'effet de facilitation observé entre le mot dérivé et sa racine n'est pas réductible à un effet purement phonologique. En effet, pour les paires phonologiques de type "alarme"-"larme" ou "louper"-"loup" on n'observe aucun effet de facilitation.

2- Pour les deux types de mots dérivés, on observe un effet d'amorçage entre deux mots de même nature (préfixé-préfixé; suffixé-suffixé). Cet effet confirme l'existence de liens facilitateurs entre ces mots. Ces liens pourraient s'établir via la racine.

3- On n'observe aucun effet significatif d'amorçage purement phonologique ni pour les mots préfixés ni pour les mots suffixés.

4- Pour les mots sémantiquement liés on confirme l'effet classique de facilitation sémantique. Cependant, cet effet ne remet pas en cause la nature morphologique des autres effets: si les effets étaient dus à la liaison sémantique entre les membres de chaque couple, on devrait observer un effet d'amorçage dans la condition de liaison suffixé-racine phonologiquement opaque mais sémantiquement proche, or aucun effet n'émerge.

La principale conclusion qu'il est possible de dégager de cette expérience est que l'effet de facilitation morphologique dans une situation d'amorçage intermodal ne peut pas être assimilé à un pur effet de facilitation formelle de nature phonologique, ce résultat est conforme à celui obtenu récemment par Grainger, Colé et Ségui (1991) et Drews et Zwitserlood (1995).

Nous pensons que l'explication de ces résultats et, en particulier, de la différence de traitement suivant les différents types de mots, réside dans la structure même de la langue française.

En français les mots préfixés (et plus particulièrement ceux que nous avons utilisés) sont transparents morphologiquement et sémantiquement. La plupart des préfixes ont leur propre sens comme "re" = "à nouveau" ou "in"= "pas". Ce qui signifie que lorsqu'on dérive une racine en lui ajoutant un préfixe, on ajoute un sens au mot mais on ne change pas la catégorie grammaticale de l'item.

En revanche les suffixes n'ont pas un sens évident comme "...tion" ou "...er", la fonction du suffixe consiste essentiellement à changer la catégorie syntaxique de la racine.

On aboutit à la conclusion que pour les mots préfixés la structure morphologique est tellement prégnante, qu'il y aurait une décomposition post lexicale irrépressible et automatique, ce qui expliquerait le fait que même dans le cas où la racine n'a pas la même réalisation phonologique lorsqu'elle est libre ou intégrée, un mot préfixé active toujours sa racine.

Pour les mots suffixés, l'entrée dans le mot s'effectuant par la racine, on aurait une représentation et une entrée lexicale par la racine mais uniquement pour les formes régulières. Les formes irrégulières auraient leur propre entrée indépendante.

4. BIBLIOGRAPHIE

Beauvillain, C. & Segui, J. (1992). Representation and processing of morphological information. In R. Frost and L. Katz (Eds), Orthography, Phonology, Morphology, and Meaning. Elsevier Science Publishers B.V.

Cole, P., Beauvillain, C., & Segui, J. (1989). On the representation and processing of prefixed and suffixed derived words: A differential frequence effect. *Journal of Memory and Language*, 28, 1 - 13.

Connine C.M., Mullenix J., Shernoff E. & Yelen J. (1990). Word familiarity and frequency in visual and auditory word recognition. *Learning, Memory, and Cognition*, 16 (6), 1084-1096.

Drews, E. & Zwitserlood, P. (1995). Morphological and Orthographic similarity in visual word recognition. *Journal of experimental psychology: Human perception and performance*, 21(5), 1098-1116.

Graiger, J., Cole, P.& Segui, J. (1991). Masked morphological priming in visual word recognition. *Journal of Memory and Language*. 30. 370-384.

Mc Quenn, J.M. & Cutler, A.(1994) Morphology in Word Recognition. Manuscrit non publié.

Pillon, A. (1993). *La mémoire des mots: ses unités, son organisation*. Mardaga Psychologie et sciences humaines.

Taft, M., & Forster, K.I. (1975). Lexical storage and retrieval of prefixed words. *Journal of Verbal Learning and Verbal Behavior* , 14, 638 - 647.

Tyler, T.K., Waksler, R.& Marslen - Wilson, W.D. (1993). Representation and access of derived words in English. In Altmann, K., Schillcock (Eds), Cognitive Models of Speech Processing. The Sperlonga Meeting II, Hove, Lawrence Erlbaum Associates Ltd.

SÉGRÉGATION DE VOYELLES SIMULTANÉES: EFFETS DU NIVEAU RELATIF ET DE LA DIFFÉRENCE DE F_0

Alain de CHEVEIGNÉ

Laboratoire de Linguistique Formelle, CNRS/Université Paris 7, 2 place Jussieu, 75251, Paris

Tél.: +33 1 44273633, email: alain@linguist.jussieu.fr

ABSTRACT

Subjects were presented with stimuli consisting of single or "double" (concurrent) vowels. They had to decide whether each stimulus contained one or two vowels, and which vowels they were. In the case of double vowels, constituents had either the same fundamental frequency (F_0) or F_0s differing by 6%. They had either the same RMS level, or levels differing by 10 or 20 dB. The average number of vowels reported and the identification rate of each constituent were measured for each condition. When F_0 differed by 6%, subjects answered two vowels more often than at unison. Identification accuracy was also better than at unison when the target was at 0, -10 or -20 dB relative to the competing vowel. For stronger targets (10 or 20 dB), identification was almost perfect and therefore little affected by F_0 differences. These results are compatible with the hypothesis that segregation occurs according to a mechanism of *harmonic cancellation* rather than harmonic enhancement. They are incompatible with recent models of vowel segregation.

1. INTRODUCTION

Pour séparer perceptivement les sons, le système auditif utilise, entre autres indices, les différences de fréquence fondamentale (ΔF_0). Brokx et Nooteboom (1982) ont trouvé que l'identification de phrases concurrentes (naturelles ou synthétiques) était meilleure lorsqu'elles étaient prononcées sur des tons différents qu'à l'unisson. Le même avantage a été trouvé pour l'identification de paires de voyelles synthétiques stationnaires (Scheffers, 1983; Assmann & Summerfield 1990; Culling & Darwin 1993). De nombreux modèles et méthodes de séparation de sons harmoniques ont été proposés pour expliquer ce phénomène, ou pour le reproduire dans des systèmes de traitement de la parole (voir de Cheveigné 1993 pour une revue).

Parmi eux, le modèle de Meddis et Hewitt (1992) est accepté comme le plus plausible. Ce modèle comprend un banc de filtres dont chaque canal est traité par un modèle de cellule ciliée produisant une probabilité de décharge nerveuse. La fonction d'autocorrélation (ACF) de cette probabilité est calculée, et les ACF de tous les canaux sont additionnées pour former une ACF globale. Le maximum de l'ACF globale (dans une gamme de délais) sert à définir la *période dominante* de la réponse. Le modèle effectue alors une partition des canaux entre ceux qui sont dominés par cette période (c.a.d dont l'ACF y présente un pic), et les autres. La somme des ACF du premier groupe de canaux représente la "voyelle dominante", et la somme des ACF des canaux restants représente la "voyelle dominée". Après cette étape de ségrégation, chaque voyelle est identifiées par comparaison de la partie de l'ACF globale comprise entre 0 et 4,5 ms à des patrons de référence.

Le fonctionnement du modèle de Meddis et Hewitt dépend ainsi d'une partition de la population de canaux selon leur périodicité. Si tous étaient dominés par une seule période (par exemple si une voyelle était plus intense que l'autre), le modèle ne fonctionnerait pas et on ne constaterait alors aucun effet de ΔF_0. L'expérience décrite ici explore cette possibilité en faisant varier le niveau relatif entre voyelles. Un autre objectif était

de trouver un niveau relatif qui permette de s'affranchir des effets de plafond souvent constatés dans les expériences de "voyelles doubles": Lorsque leur niveau est égal, l'identification des deux voyelles est souvent quasi-parfaite, et donc insensible aux paramètres d'expérience. Le déséquilibre de niveau rend la tâche plus difficile pour une des voyelles, et améliore ainsi la sensibilité.

2. MÉTHODES

Des voyelles japonaises isolées ou concurrentes (somme de deux voyelles) furent présentées à des sujets japonais qui devaient juger à chaque fois s'ils entendaient une ou deux voyelles, et lesquelles. Les six sujets étaient informés que chaque stimulus comprenait une ou deux voyelles distinctes, mais ne recevaient aucun feedback.

Les cinq voyelles (/a/, /e/, /i/, /o/, /u/) furent synthétisées à une fréquence d'échantillonage de 16 kHz avec des F_0 de 125 et 132,5 Hz. Les voyelles doubles furent obtenues en formant toutes les combinaisons de F_0 (pour produire deux ΔF_0: 0 et 6%) et de voyelles distinctes (10 paires), et en additionnant les constituants avec des niveaux RMS relatifs de -20, -10, 0, 10 et 20 dB. Les voyelles doubles étaient ainsi au nombre de (10 paires) x (2 ΔF_0) x (2 ordres de F_0) x (5 niveaux relatifs) x (3 répétions) = 600 paires, auxquelles furent ajoutés 240 voyelles simples, pour un total de 840 stimuli par session. Les voyelles simples servaient à rendre l'ensemble conforme à la description faite aux sujets, et à repérer d'éventuels effets des paramètres de synthèse sur la qualité des voyelles. Après ajustement à un niveau RMS standard, les stimuli furent présentées aux sujets via casque, à un niveau sonore compris entre 63 et 70 dBA. Chaque sujet participa à cinq sessions.

Pour toutes les conditions, le nombre moyen de réponses par stimulus et le taux d'identificaton furent calculés. Chaque voyelle isolée fut jugée correctement identifiée si la réponse (une ou deux voyelles) comprenait le nom de cette voyelle. Chaque constituant de voyelle double fut jugée correctement identifiée si la réponse (une ou deux voyelles) comprenait le nom du constituant. Ces réponses furent classées selon la nature du constituant (phonème, F_0), la nature de la voyelle concurrente, et leur relation (ΔF_0, niveau relatif). Cette technique d'analyse en terme de taux d'identification des *constituants* diffère de celle utilisée dans les expériences classiques, qui mesurent généralement un taux d'identification de *paires* (deux voyelles correctes).

3. RÉSULTATS

Les résultats présentés ici sont moyennés sur les facteurs *sujet*, *session*, F_0, et *paire*. À l'unisson (Fig. 1, trait épais), les sujets tendent a faire une réponse double lorsque les voyelles sont de même niveau. Lorsque l'une domine l'autre, les réponses doubles sont plus rares, mais leur nombre reste appréciable même pour les voyelles isolées (à droite). À ΔF_0=6% (trait fin), les réponses doubles sont plus nombreuses quel que soit le niveau.

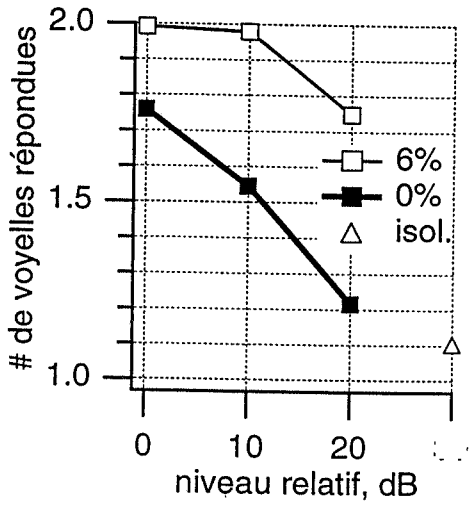


Fig. 1 *Nombre moyen de voyelles répondues par stimulus en fonction du niveau relatif, pour deux valeurs du ΔF_0. Triangle à droite: voyelles isolées.*

À l'unisson, l'identification d'une voyelle est d'autant meilleure qu'elle domine en niveau sa concurrente (Fig. 2, trait épais). À ΔF_0=6% (trait fin), le taux d'identification est plus élevé qu'à l'unisson, en particulier lorsque la cible est faible par rapport à la voyelle concurrente. L'effet de ΔF_0 est plus fort à -10 dB qu'aux autres niveaux. L'obtention d'effets expérimentaux relativement

forts est d'un intérêt pratique, puisqu'ils sont plus faciles à mettre en évidence de façon statistiquement robuste. Cet avantage serait cependant nul si la *variabilité* était également plus élevée.

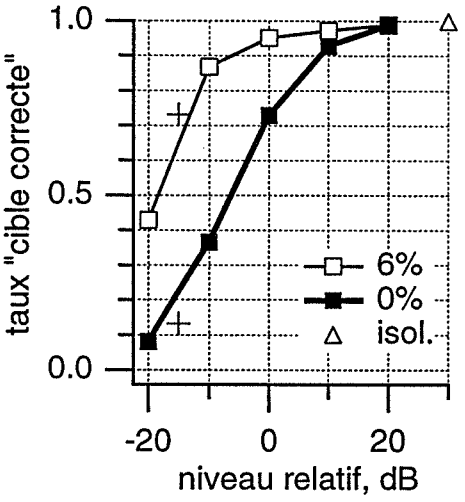


Fig. 2 *Taux d'identification d'une voyelle cible en fonction du niveau relatif à la voyelle concurrente, pour deux valeurs de ΔF_0. Croix: taux obtenus à -15 dB dans une autre expérience avec les mêmes sujets (de Cheveigné et al. 1996a). Triangle à droite: voyelles isolées.*

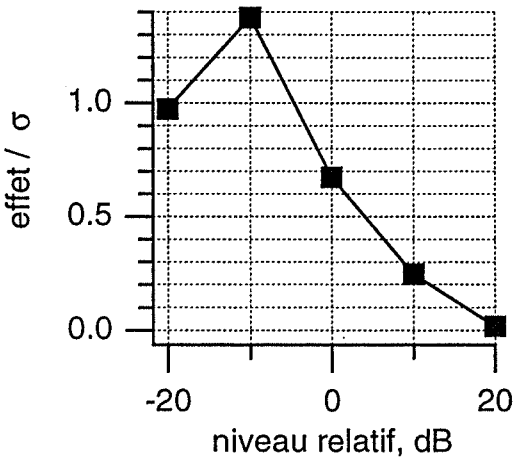


Fig. 3 *Rapport entre la taille de l'effet de ΔF_0 et sa déviation standard, en fonction du niveau relatif.*

La Fig. 3 montre qu'il n'en est rien: le rapport de la taille de l'effet (différence des taux d'identification à 6 et 0 %) à sa déviation standard est bien plus élevé à -10 dB qu'à d'autres valeurs du niveau relatif. À l'inverse un effet de ΔF_0sur une cible dominante sera difficile a mettre en evidence de façon statististiquement robuste.

4. DISCUSSION

Nos sujets pouvaient répondre une ou deux voyelles par stimulus (à la différence des expériences classiques citées en Introduction, où le sujet devait obligatoirement répondre deux voyelles). Le nombre moyen de réponses est une mesure intéressante, reflétant une perception de la "multiplicité" des sources. À l'unisson ce nombre est élevé lorsque les voyelles sont de même niveau (et donc que le stimulus ne ressemble à aucune voyelle simple). Il est également élevé à tous niveaux lorsqu'il y a un ΔF_0. Notre nouvelle procédure affecte aussi les taux d'identification, et tend à produire des effets de ΔF_0 plus marqués que dans les expériences classiques (de Cheveigné et al. 1996a).

L'identification dépend peu de ΔF_0 lorsque la cible est forte (10 ou 20 dB), du fait de l'effet de plafond. En revanche la dépendance est marquée lorsque la cible est faible (-10 ou -20 dB). Ce résultat suggère que la ségrégation s'opère selon un mécanisme d'*annulation harmonique*, selon lequel la structure harmonique (F_0) de la voyelle concurrente facilite son élimination (de Cheveigné 1993; de Cheveigné et al. 1995). En effet, le F_0 de la voyelle concurrente est facile à estimer lorsque la cible est faible. Cette hypothèse est confortée par d'autres résultats (Lea 1992; Summerfield & Culling 1992; de Cheveigné et al. 1996b). Une hypothèse rivale est que la structure harmonique de la cible faciliterait son identification (hypothèse de *renforcement harmonique*). Nous pouvons l'écarter dans les conditions où la cible est faible (-10 ou -20 dB), puisque son F_0 serait alors particulièrement difficile à estimer. Lorsque la cible est forte, l'effet de plafond nous empêche de conclure à un quelconque effet de renforcement harmonique puisque l'identification est déja parfaite sans le concours du ΔF_0. Jusqu'à présent, peu de résultats expérimentaux confortent l'hypothèse de renforcement harmonique, qui pourtant inspire nombre de méthodes d'élimination de voix parasites...

Dans l'introduction, nous avons suggéré que le modèle de ségrégation de voyelles de

Meddis et Hewitt (1992) pourrait ne pas fonctionner si une voyelle était trop dominante. Le modèle fut implémenté et appliqué à nos stimulis dans le cas ΔF_0=6%. Pour des niveaux relatifs modérés (0 ou 10 dB), la partition des canaux s'effectue bien comme prévu. En revanche à 20 dB de différence de niveau inter-voyelles, et pour certaines paires, *tous* les canaux sont dominés par la même périodicité, empêchant ainsi toute partition. Faute de partition, on ne devrait constater aucun effet bénéfique de ΔF_0sur l'identification. Pourtant nos résultats ont révélé un effet marqué pour ces mêmes paires, que le modèle de Meddis et Hewitt ne peut donc pas expliquer. Un modèle pouvant les expliquer est proposé dans de Cheveigné (1996).

5. CONCLUSION

L'identification d'une voyelle synthétique accompagnée d'une voyelle concurrente est meilleure lorsque leurs F_0 diffèrent de 6% plutôt que 0%. L'effet est le plus robuste lorsque le niveau de la voyelle cible est de -10 dB par rapport à la voyelle concurrente, et reste marqué à -20 dB, résultat que le modèle accepté comme le plus plausible, celui de Meddis et Hewitt (1992), ne peut pas prédire.

Le nombre moyen de voyelles répondues par stimulus est plus élévé lorsque les F_0 sont différents. La différence de F_0 joue le rôle d'un indicateur de *multiplicité* des sources.

6. REMERCIEMENTS

Ce travail a été conduit dans le cadre d'un accord de collaboration entre ATR Human Information Processing Laboratories, le CNRS, et l'Université Paris 7. Merci à ATR pour son hospitalité, et au CNRS pour l'autorisation d'absence. S. McAdams et C. Marin ont participé à la préparation des expériences, H. Kawahara, M. Tsuzaki et K. Aikawa ont aidé à leur élaboration, R. Kubo a assisté à leur conduite, et J. Culling a fourni les programmes de synthèse.

7. BIBLIOGRAPHIE

Assmann, P. F. and Summerfield, Q. (1990). "Modeling the perception of concurrent vowels: Vowels with different fundamental frequencies." *J. Acoust. Soc. Am.*, 88, 680-697.

Brokx, J. P. L. and Nooteboom, S. G. (1982). "Intonation and the perceptual separation of simultaneous voices," *Journal of Phonetics* 10, 23-36.

Culling, J. F. and Darwin, C. J. (1993). "Perceptual separation of simultaneous vowels: Within and across-formant grouping by F0," *J. Acoust. Soc. Am.*, 93, 3454-3467.

de Cheveigné, A. (1993). "Separation of concurrent harmonic sounds: Fundamental frequency estimation and a time-domain cancellation model of auditory processing," *J. Acoust. Soc. Am.*, 93, 3271-3290.

de Cheveigné, A. (1996). "Concurrent vowel segregation III: a neural time-domain model of harmonic interference cancellation," *J. Acoust. Soc. Am.*, en préparation.

de Cheveigné, A., McAdams, S., Laroche, J. and Rosenberg, M. (1995). "Identification of concurrent harmonic and inharmonic vowels: A test of the theory of harmonic cancellation and enhancement," *J. Acoust. Soc. Am.*, 97, 3736-3748.

de Cheveigné, A, Kawahara, H, Tsuzaki, M. and Aikawa, K. (1996a). "Concurrent vowel segregation I: effects of relative level and F0 difference," *J. Acoust. Soc. Am.*, en préparation.

de Cheveigné, A., McAdams, S., Marin, M. (1996b). "Concurrent vowel segregation II: effects of phase, harmonicity and task," *J. Acoust. Soc. Am.*, en préparation.

Lea, A. (1992). "*Auditory models of vowel perception*," unpublished doctoral dissertation, Nottingham.

Meddis, R. and Hewitt, M. J. (1992). "Modeling the identification of concurrent vowels with different fundamental frequencies," *J. Acoust. Soc. Am.*, 91, 233-245.

Scheffers, M. T. M. (1983). "*Sifting vowels*," unpublished doctoral thesis, Gröningen.

Summerfield, Q. and Culling, J. F. (1992). "Periodicity of maskers not targets determines ease of perceptual segregation using differences in fundamental frequency". 124th meeting of the ASA [*J. Acoust. Soc. Am.* 92, 2317 (A)].

TAILLE DES FENETRES PERCEPTIVES, EMPAN DE LA MEMOIRE AUDITIVE

Claire GERARD, Nina DOLGËR

Laboratoire de Psychologie Expérimentale, Université René Descartes, URA CNRS 316

Tel: 40 51 98 50 - Fax: 40 51 70 85

ABSTRACT

When silences of various durations are artificially inserted in natural sentences, the perception of an accentuated word can be used to study how long prosodic parameters are stored in the auditory short term memory. The number of correct responses first decreases and then improves again when the duration of inserted silences increases up to 400 ms. These data should call into question the values usually considered as limiting auditory storage in speech perception.

1. INTRODUCTION

Une des façons de s'assurer que les qualités d'un son (comme la sonie ou la hauteur) sont encore "présentes" dans la mémoire auditive des sujets consiste à s'appuyer sur les phénomènes de masquage pro et rétro-actifs : si deux sons sont séparés par un intervalle bref, l'un (le son masquant) empêche le traitement adéquat des caractéristiques acoustiques de l'autre (le son masqué), dont la sonie, la hauteur et/ou le timbre se trouvent modifiés. Au-delà d'un intervalle de 250 millisecondes environ, le masquage n'opère plus (Zwicker, 1982; Botte, 1989). Ceci laisserait penser que la durée de vie de la mémoire auditive exhaustive est fort courte. A propos des sons spécifiques de la parole, Massaro (1974), Sorin (1989) évaluent à 2 ou 3 syllabes (donc à environ une seconde) la durée du maintien en mémoire des paramètres acoustiques. Mais Semal (1991) rappelle que la trace "prélinguistique" laissée par un stimulus non verbal ne s'estompe pas aussi rapidement que les modèles structuraux de la mémoire pourraient le laisser croire, et les travaux de Deutsch (1970, 1975, 1982) contribuent à faire penser que les attributs sensoriels d'un stimulus coexistent en mémoire avec ses attributs sémantiques.

Qu'en est-il maintenant des traitements prosodiques portant sur de longues portions d'énoncés ? La prosodie de la parole se manifeste par des pauses, des accentuations, des variations de rythme, des changements de contour mélodique... qui sont des témoins aussi bien de l'organisation syntaxique de l'énoncé que de sa composition lexicale et de sa signification générale. Pour saisir ces indices prosodiques, et leur attribuer la signification adéquate, l'auditeur doit ignorer certaines variations sonores considérées comme sans signification propre, mais il lui faut aussi prendre en considération certaines autres variations considérées comme pertinentes et les interpréter. Cette deuxième étape implique des jugements comparatifs ("plus fort"... "plus lent"...) entre deux portions au moins de l'énoncé, - très probablement de la taille d'un mot -, donc exige que ces portions de l'énoncé soient stockées en mémoire pendant une durée que l'on peut estimer au moins égale à une seconde, puis remplacées par d'autres portions du flux sonore ultérieur, en temps réel, pendant le déroulement de l'énoncé. Ainsi, comme Sorin (1989) en fait l'hypothèse, l'auditeur ouvrirait des fenêtres perceptives successives sur la chaîne sonore et stockerait en mémoire leur contenu. L'écoute de la parole continue relève de processus de perception, de mémorisation et d'interprétation eux-mêmes continus, ce qui oblige à considérer sur le plan théorique que des intersections sont nécessaires entre fenêtres pour étendre les comparaisons "intra-fenêtres" à des comparaisons "inter-fenêtres", afin d'assurer la propagation de ces processus à toute la phrase. Quelle est la taille de ces fenêtres perceptives, en d'autres termes, quelle est la durée du flux sonore stocké en mémoire auditive lors de la perception de la parole continue ?

La perception d'un mot accentué, après altération artificielle d'énoncés naturels par insertion de pauses, peut servir de moyen d'étude de la durée du flux sonore susceptible d'être captée dans une mémoire auditive à très court terme. Notre principe expérimental est le suivant: soient deux énoncés identiques, l'un comportant un accent d'insistance et l'autre non, nous recherchons jusqu'à quel degré d'étirement dans le temps l'auditeur reste en mesure de percevoir la présence ou l'absence de l'accent. Mais la seule insertion de pauses avant et après un mot risque de promouvoir celui-ci au rang de mot accentué. Il était donc nécessaire de faire précéder l'étude principale d'expériences préliminaires.

2. METHODE

Le matériel utilisé est constitué d'un ensemble de sept phrases, chacune énoncée avec et sans accent d'insistance, extraites d'un corpus qui avait fait l'objet d'études préalables (Dahan, 1994). L'ensemble de ce corpus a été présenté à 92 auditeurs, chargés de déterminer si les phrases contenaient un accent d'insistance. Les phrases sélectionnées pour notre expérience ont été retenues 1) parce qu'elles donnaient lieu à un faible pourcentage de détection d'accent (moins de 25 %) dans leur version dite "neutre" (ne comportant pas d'accent), et à un fort pourcentage de détection d'accent dans leur version accentuée (plus de 80 %); 2) parce que les mots cibles de ces phrases ne comportaient pas d'occlusive en début de mot, ce qui permet d'obtenir une meilleure précision dans l'estimation de la durée des pauses qui les précèdent sans être gêné par la présence d'un silence articulatoire. On notera dès maintenant que les phrases naturelles énoncées sans accent que nous avons retenues donnent cependant lieu (dans la limite de 25 %) à la perception d'un accent, sans que l'analyse acoustique du signal sonore permette de mettre en évidence des variations prosodiques d'importance. Ces phrases sont les suivantes:

Les scores de détection d'un accent, pour la version accentuée des phrases, tous supérieurs à 80%, sont présentés entre parenthèses.

Phrase 1 : J'attends un appel **urgent** de l'étranger ce matin. (score de détection d'un accent: 100%)

Phrase 2 : Ce magasin solde toutes les chaussures **fermées** de la collection automne/hiver. (83%)

Phrase 3 : Je me suis acheté une **chemise** en soldes aux galeries Lafayette. (100%)

Phrase 4 : Elle m'a dit qu'elle partait en **vacances** une semaine début Février. (94%)

Phrase 5 : L'arrivée du courrier est **vraiment** très lente ces derniers temps. (95%)

Phrase 6 : Je crois qu'il téléphonera **sûrement** avant de venir ici. (97%)

Phrase 7 : Il est tombé un brouillard **givrant** sur toute la France aujourd'hui. (100%)

(Le mot accentué est écrit en caractères gras)

Ensuite, grâce à une procédure "d'échange d'organisation temporelle", réalisée à l'aide du logiciel UNICE, nous avons essayé d'appréhender l'importance des pauses dans la détection de cet accent selon 3 procédures successives.

Expérience 1: les pauses "normales" (dans la version neutre) entourant le mot cible, ont été remplacées par les pauses qui apparaissent autour du même mot cible, une fois celui-ci accentué. A l'inverse, les pauses entourant le mot accentué ont été ramenées à des valeurs inférieures, identiques à celles des pauses adjacentes à ce même mot cible énoncé sous version neutre par le même locuteur. Notre raisonnement était le suivant : si les pauses, entourant le mot cible sont suffisantes pour la détection d'accent dans les versions accentuées, leur présence autour d'un mot neutre pourrait donner l'illusion d'un accent, et la diminution de leur durée autour d'un mot accentué pourrait réduire, sinon supprimer l'impression d'accentuation.

Le résultat de cette expérience, où les autres indices prosodiques de l'accent n'ont pas été modifiés, ne permet pas d'affirmer un tel rôle (nécessaire et suffisant) des pauses: la seule présence de pauses dans des phrases au départ neutres n'ont pas conduit les 40 auditeurs pris comme sujets à la détection d'un accent (le score de détection d'un accent est resté de l'ordre de 25 %). De même, la diminution de durée des pauses entourant les cibles dans les versions au départ accentuées ne supprimait pas la perception de l'accent (le score de détection est resté supérieur à 90 %).

Expérience 2: Ces résultats auraient pu aussi provenir du fait que les pauses ainsi "échangées" n'avaient pas des valeurs assez différenciées. En effet, les valeurs des pauses spontanées entourant les mots cibles dans les phrases neutres s'étageaient de 20 à 70 ms, et leur valeur moyenne était de 35 ms. Ces mêmes pauses, dans les versions accentuées, fluctuaient entre 60 et 90 ms, avec une valeur moyenne de 80 ms. Les valeurs minimales des pauses effectuées dans les versions accentuées pouvaient donc être proches des valeurs maximales des pauses enregistrées dans les versions neutres. Dans un deuxième temps, les pauses artificielles insérées autour des mots "neutres" ont alors été augmentées de 150 à 200 ms selon leur durée initiale. Le matériel de la seconde expérience a été réduit aux phrases neutres, étant donné que l'effet d'accentuation des mots réellement accentués n'était pas modifiable par suppression des pauses, comme l'avait montré la première expérience. Mais cette fois encore, les mots neutres entourés de ces nouvelles pauses n'étaient pas perçus majoritairement comme accentués par 20 nouveaux auditeurs, le score de détection étant resté de 26%.

Expérience 3: Une troisième pré-expérience nous a alors été inspirée par les travaux de Butcher (1975) qui a montré que, pour être détectée *en tant que telle* au moins 75 % des fois, une pause doit atteindre des valeurs différentes selon son emplacement dans la

phrase : une pause située en frontière de phrase doit être très supérieure (plus de 1000 ms) à une pause située entre constituants principaux (entre 500 et 1000 ms) pour être détectée. Les pauses situées au sein des constituants sont perçues pour des valeurs plus faibles (de l'ordre de 200 ms). A la lumière de telles conclusions, une détection d'*accent* pouvait être possible pour des mots au départ neutres, du moment qu'on variait la durée des pauses artificielles en fonction de la position des cibles par rapport aux constituants. Le matériel était toujours constitué d'énoncés naturels neutres uniquement, comportant des pauses artificielles correspondant aux valeurs indiquées par Butcher (op. cit.). Cependant, même avec de telles durées de pauses autour des mots cibles, les sujets (20 auditeurs différents des précédents) ne considéraient toujours pas le mot neutre comme accentué de façon majoritaire (score = 28 %).

Ces expériences nous ont donc permis de montrer que, pour le matériel spécifique que nous avons sélectionné, la seule insertion de pauses, même longues, ne suffit pas à promouvoir un mot neutre au rang de mot accentué, et inversement, que la totale suppression de pauses ne suffit pas à rendre neutre un mot au départ accentué. Nous pouvions alors nous livrer à l'expérience principale, qui va consister à étirer dans le temps l'intégralité d'un énoncé. En effet, l'accent d'insistance "contamine" des portions d'énoncé éloignées de la cible, (Gérard & Dahan, 1995). Il est alors légitime d'envisager de modifier l'ensemble de l'organisation temporelle de la phrase dans le but de tester jusqu'à quel degré "d'étirement dans le temps" la perception d'un accent d'insistance naturel subsiste.

3. EXPERIENCE PRINCIPALE.

3.1 METHODE

Nous avons inséré des silences de durées égales entre chacun des mots des 7 phrases. Cette procédure altère l'intelligibilité des phrases en détruisant la coarticulation, et rend donc l'ensemble de la tâche plus difficile, mais notre seul but était d'étudier, toutes choses égales par ailleurs, l'effet des variations de la durée des silences insérés. Pour chaque phrase, quatre versions ont été construites en fonction de la durée D (en millisecondes) des silences insérés: D1=100, D2=200, D3=300, et D4=400 ms. Les mêmes énoncés ont également été synthétisés par la carte et le logiciel de synthèse PSOLA (CNET), sous une forme "neutre" (sans accent), et les mêmes durées de silences ont été insérées. Cette procédure nous a semblé être la seule qui garantisse l'absence réelle d'accentuation, alors que les énoncés naturels "neutres" ne sont peut-être pas si "neutres" que ça, à tout le moins si l'on considère que les 25 % de détection d'accent proviennent de modifications acoustiques réelles bien que non décelables aisément par l'analyse acoustique. Le corpus ainsi construit comportait alors 56 phrases correspondant au croisement de trois facteurs : 7 contenus sémantiques X 4 durées de silence X 2 formes (accentuée et non accentuée).

Les phrases naturelles et artificielles ont ensuite été présentées à 30 auditeurs (distincts des précédents) qui avaient toujours comme consigne d'indiquer s'ils percevaient ou non un mot accentué, mais qui devaient aussi noter ce mot le cas échéant, et évaluer (dans tous les cas) leur degré de certitude dans leur réponse sur une échelle de 1 à 5 points. Le rythme de présentation des 56 phrases était soutenu : 15 secondes étaient laissées entre les phrases pour donner les réponses. Cinq types de réponses, chacun assorti d'un indice de certitude, sont alors possibles:

- Phrase sans accent : réponse "pas d'accent" = ***bonne réponse.***
- Phrase sans accent : accent détecté à tort = ***fausse alarme.***
- Phrase accentuée : accent bien localisé = ***bonne détection.***
- Phrase accentuée : accent non détecté = ***oubli.***
- Phrase accentuée : accent mal localisé = ***erreur de localisation.***

Les indices de certitude et les divers types de réponse ont fait l'objet d'analyses statistiques.

3.2 RESULTATS

Deux types de réponses correctes nous intéressent : les bonnes réponses pour les phrases non accentuées, et les bonnes détections pour les phrases accentuées, dont nous allons examiner l'évolution en fonction de l'étirement de l'énoncé dans le temps. Dans un premier temps, pour les phrases accentuées comme pour les phrases non accentuées, on observe une diminution des scores de réponse correcte (bonne détection et bonne réponse), diminution qui n'est significative que pour les phrases accentuées $F(1,29) = 15,03$ $p<.001$) lorsque la durée des silences insérés artificiellement passe de 100 à 200 ms. Cette chute est suivie d'une stagnation (pour 300 ms d'intervalle entre les mots), et enfin d'une remontée lorsque la durée du silence passe de

300 à 400 ms. Cette remontée est significative pour les deux types de phrases ($F(1,29) = 9,16$ $p<.01$ pour les bonnes détections des phrases accentuées, et $F(1,29) = 4,34$ $p<.05$ pour les bonnes réponses dans le cas des phrases sans accent). Les scores correspondant aux deux types de réponses correctes s'organisent donc selon une courbe en U quand l'étirement dans le temps de l'énoncé augmente, et il faut souligner que la remontée finale de la courbe s'observe pour chacune des 7 phrases dans leur version neutre, et pour 6 phrases sur les 7 dans la version accentuée. Les valeurs des indices de certitude associés aux réponses correctes sont plus élevées que celles associées aux erreurs : en moyenne 3,3 sur l'échelle de 1 à 5 pour les réponses correctes, et 2,7 pour les oublis, fausses alarmes et erreurs de localisation ($F(1,25)= 15,9$ $p<.0005$). Mais ces indices de certitude n'évoluent pas en fonction de l'étirement de l'énoncé dans le temps, donc ne suivent pas la courbe en U des scores de réponse correcte. La figure 1 présente cette courbe en U.

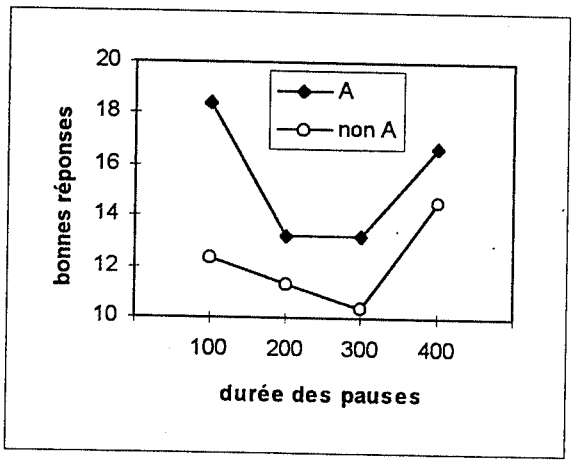


Figure 1: Evolution du nombre moyen de bonnes détections d'un accent pour les phrases accentuées (codées A) et des bonnes réponses pour les phrases neutres (codées non A) en fonction de la durée des pauses insérées entre chaque mot des phrases. Le score maximum est de 30.

4. DISCUSSION

Pour un psychologue, une telle organisation des données évoque soit un changement de stratégie perceptive en fonction de la durée des silences insérés, soit une durée de stockage plus longue que traditionnellement admis concernant la mémoire auditive, soit encore de nouveaux phénomènes de codage en mémoire. Huggins (1975) a travaillé sur les stratégies de perception de la parole dans une tâche de *shadowing*. Les segments de parole que les sujets devaient répéter au fur et à mesure de leur écoute variaient de 30 à 200 ms d'une séquence sonore à une autre, et la durée des silences insérés entre segments dans une même séquence variait de 10 à 1000 ms. Huggins observait aussi une courbe en U des performances, et interprétait ses données en invoquant une stratégie de *gap-bridging*: les auditeurs "enjamberaient" (ignoreraient) les silences lorsque ceux-ci atteignent une valeur critique propre à cette tâche. Il n'est pas exclu qu'un tel phénomène se produise dans notre expérience, mais l'explication ne semble pas suffisante. En effet, il faut bien que certaines des caractéristiques acoustiques de l'accent, autres que les pauses, restent stockées en mémoire auditive plus longtemps qu'une seconde ou deux pour que nos sujets augmentent leurs scores de détection de l'accent lorsque les pauses atteignent 400 millisecondes. Dans notre situation, les phrases s'étirent au minimum sur 3 secondes (400 ms x 7 mots minimum d'une phrase, plus la durée intrinsèque de chaque mot...). Les limites en durée de stockage de la mémoire auditive devraient-elles être remises en question ?

5. BIBLIOGRAPHIE

Butcher, A. (1975). Pause and syntactic structure, in H.W. Dechert, & M. Raupach (Eds.), *Temporal variables in speech*, The Hague, Paris, New-York: Mouton, 85-90.

Botte, M.C. (1989). L'audition, in C. Bonnet, R. Ghiglione et J.F. Richard (Eds.), *Traité de Psychologie Cognitive*, Paris: Dunod, 83-127.

Dahan, D. (1994). Etude de la prosodie du français en parole continue, Thèse de Troisième Cycle, Paris V.

Deutsch, D. (1970). Tones and numbers: Specificity of interference in short-term memory, *Science*, 168, 1604-1605.

Deutsch, D. (1975). Auditory memory, *Canadian Journal of Psychology*, 29, 87-105.

Deutsch, D. (1982). The organization of short-term-memory for a single acoustic attribute, in D. Deutsch & J. Deutsch (Eds.), *Short-term memory*, New York: Academic Press, 107-151.

Gérard, C. & Dahan, D. (1995). Speech durational variations and semantic focusing in reading, *Speech Communication*, 16, 293-311.

Huggins, A.W.F. (1975). Temporally segmented speech, *Perception & Psychophysics*, 18, 226-231.

Massaro, D.W. & Cohen, M.M. (1975). Perceptual auditory storage in speech recognition, in A. Cohen & S.G. Nooteboom (Eds.), *Structure and process in speech perception*, Berlin: Springer Verlag, 226-245.

Semal, C. (1991). La mémoire auditive: une organisation modulaire ?, in J.C. Risset & G. Canévet (Eds.), *Genèse et Perception des sons*, Marseille: Publications du Laboratoire de Mécanique et d'Acoustique, 128, 205-212.

Sorin, C. (1989). Perception de la parole, in M.C. Botte, G. Canévet, L. Demany and C. Sorin (Eds.), *Psychoacoustique et Perception Auditive*, Paris: Editions INSERM/EMI, 123-139.

Zwicker, E. (1982). *Psychoakustik*, Berlin: Springer Verlag.

TRAITEMENT DES INDICES METRIQUES ET DES INDICES PHONOTACTIQUES LORS DE LA SEGMENTATION LEXICALE

Marie-Hélène BANEL , Nicole BACRI

Université René Descartes, Laboratoire de Psychologie Expérimentale, URA 316 CNRS, 28 rue Serpente, 75006 Paris.

ABSTRACT

This research investigates the contributions of metrical and phonotactic segmentation cues in French word recognition. In experiment 1, listeners detected monosyllabic words embedded in initial position in nonsense bisyllabic strings, which were realized either with the short-long usual pattern, or with a long-short or a long-long pattern. Bisyllabic strings contained either a medial illegal cluster or an 'enchaînement'. Results showed an effect of both Metrical patterns and Phonotactic cues without any interaction. Experiment 2 studied the recognition of words extracted from the bisyllabic strings. It showed no effect of previous context. Results are discussed in terms of timing mechanisms that may operate either at the syllabic level, or as one of the basis of a prosodic parsing routine. We conclude that phonotactic and metrical cues are processed separately.

1. INTRODUCTION

Plusieurs recherches récentes ont mis en évidence le rôle de la prosodie lors de la segmentation lexicale. Pour des langues à accent lexical, telles que l'anglais, les syllabes accentuées déclenchent la segmentation (Cutler & Norris, 1988; McQueen et al., 1994). En français l'accent n'a pas de valeur lexicale distinctive mais sa position est fixe, donc prévisible. Il est placé sur la syllabe finale des mots plurisyllabiques, et se marque par un allongement perceptivement saillant de cette syllabe. Malgré la tendance à accentuer la syllabe initiale des mots dans certains styles oraux (par ex. le style radiophonique, Fonàgy, 1980; Pasdeloup, 1988), la structure bref-long apparaît comme la structure de base (Vaissière, 1992). Des expériences réalisées avec des mots bisyllabiques contenant deux mots monosyllabiques ont mis en évidence une Stratégie de Segmentation Métrique en français. Un *pattern long-bref inhabituel* entrainaît plus de segmentations que le *pattern bref-long usuel.* Un effet de la structure syllabique des monosyllabes enchâssés n'etait observé que pour le pattern neutre long-long: une syllabe initiale CVC augmentait le nombre des jugements de segmentation (Banel & Bacri, 1994).

Toutefois la tâche de jugement ne permettait pas d'étudier les relations entre indices métriques et indices structuraux, à la différence de la tâche d'extraction de mots inclus dans une séquence de syllabes sans signification dont nous aller présenter les résultats. On comparera les effets de deux indices structuraux opposés: un enchaînement et une suite consonantique illégale qui introduit une rupture. La suite illégale consistera en une succession de deux consonnes qui ne peuvent former un groupement consonantique au début ou à la fin d'un mot. La confrontation directe des effets des indices phonotactiques et métriques devrait fournir des informations sur le niveau de traitement des patterns métriques en français.

2. EXPERIENCE 1

Les poids respectifs de chaque indice sont examinés lors de la comparaison de suites **CVCCVC** (*langzok*) portant un *cluster illégal* médian (Dell, 1995), et de suites **CVCVC** (*langok*). Toutes les suites débutent par un mot **CVC** (*langue*) et s'achèvent par une séquence sans signification. Dans les **CVCCVC**, la consonne finale du mot CVC est en position de *coda,* et la consonne suivante en position d'*attaque.* Dans les **CVCVC**, la consonne finale du mot est en position d'*attaque* de la seconde syllabe et est dite enchaînée. L'extraction du mot CVC exige en ce cas une resyllabation pour replacer la consonne finale en position de *coda.* Cette opération devrait augmenter le temps d'extraction du mot des CVCVC par rapport aux CVCCVC. Si les sujets appliquent une Stratégie de Segmentation Métrique, le pattern long-bref facilitera l'extraction du mot relativement à la condition contrôle, alors que le pattern bref-long la ralentira puisqu 'il suscite le traitement de la suite bisyllabique entière. Si

les deux types d'indices sont traités indépendamment, aucune interaction ne sera observée entre indices métriques et indices structuraux.

2.1. Méthode

2.1.1. Sujets

60 étudiants en licence de psychologie à l'université Paris V, tous de langue maternelle française.

2.1.2. Matériel et procédure

Les stimuli test se composaient de 18 mots monosyllabiques CVC enchâssés dans des suites CVCCVC et dans des suites CVCVC. Chacun des deux types de suites a été combiné avec les patterns métriques bref-long usuel pour les bisyllabes, long-bref inversé et long-long neutre. Pour chaque item test, six versions ont été construites, définissant six conditions expérimentales. Ces conditions résultaient du croisement entre les facteurs Type d'indice phonotactique et Pattern métrique. Les items de remplissage étaient composés de deux syllabes sans signification. Un sujet n'avait à détecter la présence d'un même mot CVC au début des séquences que dans l'une des six conditions. Après détection, le mot était prononcé à voix haute. Les sujets ont été confrontés aux six conditions expérimentales, l'ordre de présentation étant contrebalancé. Les temps de réponse (TR) ont été mesurés à partir de la fin des mots CVC.

2.2. Résultats et discussion

Deux items test qui n'ont pas été reconnus par 50 % des sujets ont été exclus des analyses. Les résultats sont présentés dans la Table 1.

Une analyse de variance a montré que l'effet principal du type d'indice phonotactique est significatif ($p<.001$). La présence d'une suite illégale a facilité la segmentation (TR moyen: 782 ms) par comparaison avec les temps obtenus en présence d'une consonne médiane perçue à l'attaque de la seconde syllabe (840 ms). L'effet principal du pattern métrique est lui aussi significatif ($p<.0001$). Le pattern métrique long-bref inhabituel pour les bisyllabes (703 ms) a facilité la segmentation par rapport au pattern neutre (840 ms; $p<.001$), et par rapport au pattern usuel bref-long (889 ms; $p<.00001$). La différence de 49 ms entre les patterns bref-long et neutre n'est pas significative. L'interaction entre le type d'indice phonotactique et le pattern métrique n'est pas significative ($F<1$).

Cette absence d'interaction correspond à une distribution des performances semblable pour les deux types d'indices phonotactiques. Pour les items CVCCVC (Table I), la présence d'un conflit entre informations phonotactiques et métriques, suscité par un pattern bref-long, a ralenti les détections par rapport à la condition contrôle, mais l'interférence entre informations que traduit ce ralentissement n'est pas significative. La possibilité d'une coopération entre indice phonotactique et pattern long-bref a facilité l'extraction du mot (gain: 117 ms, $p<.01$). La condition de coopération a entrainé des détections plus rapides que la condition de conflit: la différence entre ces deux conditions a atteint 179 ms ($p<.001$).

La constatation de gains significatifs quand le pattern est long-bref, d'interférences, pour faibles qu'elles soient, lorsqu'il est bref-long montre le rôle du traitement d'un contraste métrique dans la détection des mots. Toutefois, on pourrait faire l'hypothèse que les différences de performances sont dues à des différences acoustiques entre les classes de stimuli. Les corrélats acoustiques des frontières de syllabes sont nombreux. On sait, par exemple, que la durée d'une occlusive en position initiale (CVC.CVC) est supérieure à celle d'une occlusive en position finale (CVC.CVC) (Quené, 1993). Bien que les stimuli constituant notre matériel aient été sélectionnés de façon à minimiser ces différences, l'analyse acoustique montre par exemple que la durée du phonème /p/ en position initiale est plus longue qu'en position finale en moyenne de 25 ms. En dépit de l'absence de significativité de ces différences, liée à une forte dispersion sur l'ensemble des items, il ne peut être exclu que les sujets aient utilisé les corrélats acoustiques des frontières de syllabe et de mot au lieu d'être d'abord guidés dans leur traitement par les contraintes phonologiques.

Table 1: Temps de réponse moyens (ms) en fonction des différentes combinaisons du pattern métrique avec le type d'indice phonotactique (s= écart-types); expérience 1.

		PATTERN METRIQUE			
		Bref-long	*Long-bref*	*Long-long*	
		CONFLIT	COOPERATION	CONTROLE	moyennes
TYPE D'INDICE PHONOTACTIQUE	*CVC.CVC 'langzok'*	**862** (s = 313)	**683** (s = 249)	**800** (s = 246)	782
	CV.CVC 'langok'	**916** (s = 314)	**724** (s = 232)	**880** (s = 307)	840
	moyennes	889	704	840	

3. EXPERIENCE 2

Cette expérience a pour objectif de préciser si des corrélats acoustiques des frontières phonétiques pouvaient ou non avoir servi de base à la reconnaissance des mots. Elle permettra, en second lieu, d'étudier dans quelle mesure les variations de la durée des mots enchâssés influencent leur détection (Miller & Volaitis, 1989; Wayland, Miller & Volaitis, 1994). Une étude antérieure réalisée sur le français a montré que des variations modérées de la durée intrinsèque n'affectent pas le traitement des mots lors d'une tâche de détection d'un phonème en position initiale (Dupoux & Mehler, 1990). Il se peut que la tâche de détection d'un mot monosyllabique soit plus sensible aux variations de durée que la détection d'un phonème. La détection de mot contraint le sujet à traiter l'ensemble de la syllabe alors que la détection de phonème n'implique pas nécessairement le traitement de syllabes entières (Dupoux, 1993). L'absence d'effet sur le traitement des mots de leur contexte d'occurrence originel et de leur durée confirmera l'interprétation des expériences précédentes, alors qu'un effet significatif de ces deux facteurs l'affaiblirait fortement.

3.1. Méthode

3.1.1. Sujets

20 autres étudiants.

3.1.2. Matériel et procédure

Le matériel est constitué de l'ensemble des mots cibles CVC et des items de remplissage de l'expérience précédente, extraits de leur séquence bisyllabique d' origine. Dans chacun des deux groupes de 10 sujets, chaque sujet a entendu deux des quatre versions possibles d'un item test (un mot de durée brève extrait de l'un des contextes, et le même mot de durée longue extrait de l'autre contexte, ou inversement, l'ordre de présentation des différentes versions étant contrebalancé). La tâche des sujets était d'appuyer le plus rapidement possible sur le bouton de réponse chaque fois qu'ils entendaient un mot, puis de répéter ce mot à voix haute.

3.2. Résultats et discussion

L'effet des différents facteurs sur les temps de réponse a été peu marqué. Un mot extrait d'une suite CVCVC ('*langue*' extrait de '*langok*') n'est pas plus long à détecter (545 ms; écart-types de l'ordre de 200 ms) qu'un mot extrait d'une suite CVC CVC ('*langue*' extrait de '*langzok*') (542 ms). L'extraction des mots rares (540 ms) n'est pas plus longue que celle des mots de fréquence élevée (536 ms). Un mot bref est plus difficile à détecter (561 ms) qu'un mot long (526 ms).

L'analyse de variance confirme cette description: ni le contexte d'origine, ni la fréquence des mots n'ont eu d'effet significatif ($F<1$). L'effet de la durée est significatif dans l'analyse par sujets, mais non dans l'analyse par items ($F1\ (1, 19) = 7.7$, $p<.01$; $F2\ (1, 15) = 1.98$, $p>.10$). Aucune interaction n'est significative.

Ici, l'absence de tout effet du contexte d'origine des monosyllabes permet d'affirmer que les différences entre les temps de détection liées, dans l' expérience 1, à la présence ou à l'absence d'une suite illégale de consonnes en

position médiane peuvent être interprétées comme dues à l'effet d'une forte contrainte phonologique. Les indices phonotactiques, et non les différences acoustiques *per se,* en sont la source. Toutefois l'effet de la durée du mot sur sa reconnaissance est significatif. Il semble dû à la difficulté de traiter quelques mots. Cet effet est très loin du seuil de significativité dans l'analyse par items, et n'est donc pas généralisable. Ce résultat peut néanmoins être interprété en référence au modèle de Wayland et al. (1994). Un mécanisme de traitement du 'timing' intrinsèque des syllabes - appartenant aux processus de perception des stimuli auditifs - influencerait la perception des syllabes isolées. Un mécanisme de 'timing' extrinsèque ayant pour fonction d'extraire des informations de plus haut niveau et distribuées sur de longs fragments de parole produirait ici les informations relatives aux patterns métriques.

4. DISCUSSION GENERALE

Cette recherche met en évidence des effets conjoints des contrastes métriques et des contraintes phonotactiques lors de la segmentation lexicale en français. Ces effets conjoints ne sont pas le résultat d'interactions à des étapes précoces dans le traitement perceptif. Les résultats montrent que les indices métriques ont eu un rôle prédominant en cas de conflit avec les indices phonotactiques, au point que la situation de conflit n'a pas entrainé d'interférence si on la compare à la situation contrôle (suites portant un cluster illégal). La structure métrique activerait une procédure de segmentation spécifique dont l'effet pourrait être renforcé par celui d'une discontinuité phonotactique en situation de coopération entre indices.

Le modèle de segmentation lexicale de McQueen et al. (1994) doit être modifié pour rendre compte des résultats. La Stratégie de Segmentation Métrique du français ne s'appuie pas sur un mécanisme de détection de syllabes accentuées en tant que telles, mais sur des contrastes temporels. Cette stratégie permet d'anticiper les frontières de mot. Elle s'applique en parallèle avec le traitement des indices phonotactiques, à une étape précoce de traitement du signal.

5. BIBLIOGRAPHIE

Banel M-H. & Bacri N. (1994) 0n metrical patterns and lexical parsing in French, *Speech Communication*, n° 15, 115-126.

Cutler A. & Norris D. (1988) The role of strong syllables in segmentation for lexical access, *J. of Experimental Psychology: Human Perception and Performance*, n° 14, 113-121.

Dell F. (1995) Consonant clusters and phonological syllables in French, *Lingua*, n° 95, 5-26.

Dupoux E. (1993) The time course of prelexical processing: the syllabic hypothesis revisited, in G. Altmann et R. Shillcock (Edit.), *Cognitive Models of Speech Processing: The 2d Sperlonga meeting,* New-York, L.E.A.P., Hillsdale, 81-114.

Dupoux E. & Mehler J. (1990) Monitoring the lexicon with normal and compressed speech: frequency effects and the prelexical code, *Journal of Memory and Language*, n° 29, 316-335.

Fonàgy I. (1980) L' accent français: accent probabilitaire, in Fonàgy, I, & Léon, P. R (eds), L'accent en français contemporain. *Studia Phonetica*, n° 15, 123-233.

Mc Queen J. , Norris D. & Cutler A. (1994) Competition in spoken word recognition: Spotting words in other words, , *J. of Experimental Psychology: Learning, Memory and Cognition,* n° 20, 621-638.

Miller J. L. & Volaitis L. E. (1989) Effect of speaking rate on the perceptual structure of a phonetic category, *Perception and Psychophysics*, n° 46, 505-512.

Pasdeloup V. (1988) Essai d' analyse du système accentuel du français: Distribution de l'accent secondaire, *Actes des 17èmes Journées d'Etude sur la Parole*, Nancy, 20-23 septembre.

Quéné H. (1993) Segment durations and accent as cues to word segmentation in Dutch, *Journal of the Acoustical Society of America,* n° 94, (4), 2027-2035.

Vaissière J. (1992) Rhythm, accentuation and final lenghthening in French, in Sundberg, L., Nord, L., et Carlson, R. (eds), *Music, Language, Speech and Brain,* Wenner-Green International Symposium Series, n° 59, 108-120.

Wayland S. C., Miller J. L. & Volaitis L. E. (1994) The influence of sentential speaking rate on the internal structure of phonetic categories, *Journal of the Acoustical Society of America*, n° 95, (5), 2694-2701.

TRANSGRESSIONS PHONOTACTIQUES : LE CAS DES CLUSTERS /DL/ ET /TL/ EN POSITION INITIALE

Pierre Hallé, Juan Segui, Uli Frauenfelder, & Christine Meunier

LPE (CNRS-Paris V), Paris, et Laboratoire de Psycholinguistique, Université de Genève

ABSTRACT

We report in this study a case of perceptual assimilation, whereby French listeners hear /kl/ and /gl/ instead of, respectively, /tl/ and /dl/ in word-initial position (where /tl/ and /dl/ are impermissible). Gating experiments show that the components of the illegal clusters are correctly perceived as the intended phones /t/ + /l/ or /d/ + /l/. The perceptual shift begins to unfold after the liquid component has been identified. We propose to call this kind of perceptual shift "contextual perceptual assimilation". The last part of the paper shows that this kind of assimilation effect is strongest in on-line processing of speech and entails a temporal processing cost.

1. INTRODUCTION

Il arrive souvent que l'on entende dans le flot de parole, des sons de parole, voire des mots, qui en sont pourtant objectivement absents (e.g., phonemic restoration, Warren, 1972). On reconstitue alors inconsciemment un message sonore cohérent avec l'organisation de la langue à tous les niveaux, y compris au niveau phonologique. La même tendance à faire entrer tout matériel sonore dans le moule de la langue maternelle est également illustré par le phénomène de "surdité phonologique" —terme introduit par Léon Robel, 1969— qui fait référence aux difficultés de perception des sons d'une langue étrangère. Un cas typique est celui des locuteurs japonais qui ne peuvent distinguer "Laurent" de "Roland". Les phonèmes non natifs sont pour ainsi dire passés au filtre du système phonologique natif. Mais ceci ne vaut pas seulement pour les phonèmes isolés. Les séquences de phonèmes sont elles aussi perçues à travers le filtre de la langue maternelle. Ainsi, le Japonais interdisant les clusters en attaque de syllabe, un japonais entendra /do-ra-ma/ pour l'anglais "drama" (Polivanov, 1931, voir aussi Dupoux et al., sous presse). Les contraintes phonotactiques font donc partie du "filtre phonologique".

Dans cette étude, nous nous sommes posés la question de savoir comment peuvent être perçus, *au sein de la langue native*, les clusters phonotactiquement illégaux. Certains clusters illégaux sont peut-être perçus comme formés des éléments qui les composent effectivement : la transgression est alors "transparente" ; d'autres sont peut-être irrépressiblement "assimilés" à des clusters phonétiquement proches et légaux. Une étude de Massaro et Cohen (1983) illustre cette seconde possibilité : elle montre que la frontière catégorielle entre le /r/ et le /l/ anglais est déplacée par le contexte consonantique gauche en sorte que l'on entend plutôt /r/ que /l/ après /d/ ou /t/, mais plutôt /l/ que /r/ après /s/ : dans tous les cas, les auditeurs "préfèrent" entendre des clusters légaux comme /tr/ ou /sl/ plutôt que */tl/ ou */sr/. Les clusters /dl/ et /tl/ en position initiale seraient-ils assimilés à des clusters légaux ? À quels clusters ? L'assimilation se traduit-elle par une difficulté perceptive mesurable ? Nous tentons de répondre à ces questions dans le cas du français.

Dans les expériences qui suivent, tous les stimuli sont des non-mots, pour éviter que n'entrent en jeu des processus top-down d'origine lexicale.

2. EXPÉRIENCE PRÉLIMINAIRE

Un test de transcription libre était une approche simple pour savoir si les clusters /dl/ et /tl/ sont ou non assimilés à d'autres clusters et auxquels. Dix-huit sujets français ont participé à ce test. Leur tâche était de transcrire "naïvement" les items qui leur étaient présentés auditivement. Parmi ces items, 32 non-mots enregistrés par un locuteur français, huit étaient pertinents pour notre propos :

'tlabod', 'tlabdo', 'tlobad', 'tlobda' et leurs contreparties pour le voisement 'dlapot', 'dlapto', 'dlopat', et 'dlopta'.

Pour ces 8 items, les sujets ont donné des transcriptions correspondant la plupart du temps à /kl/ ou /gl/ (85,4%), rarement à /tl/ ou /dl/ (13,2%), encore plus rarement à /pl/ (1,4%), jamais à /tr/ ou à /dr/.

Table 1 : Transcriptions des items en /tl/ et /dl/.

items	*transcrits comme :*				
	/tl/	/kl/	/dl/	/gl/	/pl/
/tl/-	10%	82%	1%	4%	3%
/dl/-	-	4%	15%	81%	-

Le résultat très net est donc que les sujets semblent entendre une occlusive vélaire au lieu d'une dentale. Si assimilation il y a, elle ne se manifeste pas par un déplacement de la perception de la liquide de /l/ à /r/, mais par un déplacement de la perception de l'occlusive initiale. Mais s'agit-il d'assimilation perceptive ? Une explication alternative vient tout de suite à l'esprit : les items utilisés avaient en fait une qualité acoustique plutôt vélaire que dentale. Il fallait examiner ce point avant de poursuivre.

3. EXPÉRIENCES DE GATING

La technique de "dévoilement progressif", c'est à dire de "gating", (Grosjean, 1980) se présentait naturellement pour examiner si des indices d'articulation vélaire étaient ou non présents (et prégnants) dans le signal pour les items critiques présentés. Nous avons utilisé une variante du paradigme de gating, où les sujets devaient transcrire ce qu'ils entendaient (au lieu de "deviner" des mots). Une qualité intrinsèquement vélaire des items en /tl/ et /dl/, aurait dû se traduire, dès que les fragments présentés induisent une perception stable du lieu d'articulation, par une majorité de transcriptions correspondant à une initiale vélaire.

Les 8 items en /dl/ et /tl/, plus 8 items de remplissage, dont 4 items contrôle en /dr/, /br/, /gr/ et /kr/ ont été présentés de façon incrémentale. Le fragment le plus court comprenait 10 ms après le relâchement de l'occlusive initiale ; la durée des fragments suivants augmentait avec un pas de 20 ms, jusqu'au 10ème fragment comprenant 190 ms après le relâchement, soit la majeure partie de la première syllabe. Deux formats de présentation ont été utilisés : le format classique en "série" où chaque séquence comprend les fragments de durée croissante d'un même item ; le format "parallèle" où les stimuli sont groupés par durée. Les résultats obtenus avec 28 sujets de langue française pour le format série et 26 pour le format parallèle étaient similaires. Les Figures 1 et 2 montrent ceux obtenus pour le format parallèle.

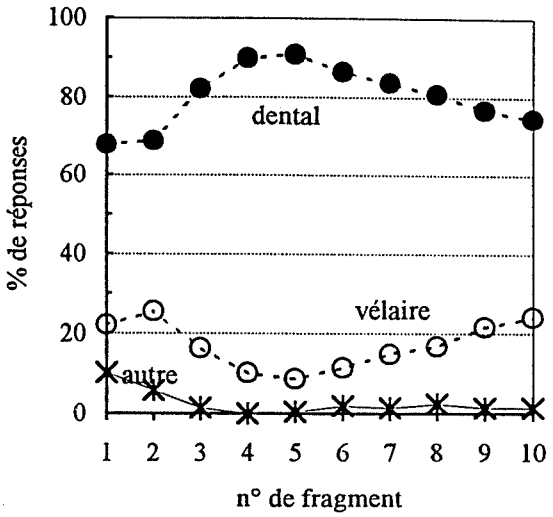


Figure 1 : Jugements dental/vélaire selon le fragment (items en /tl/ et /dl/).

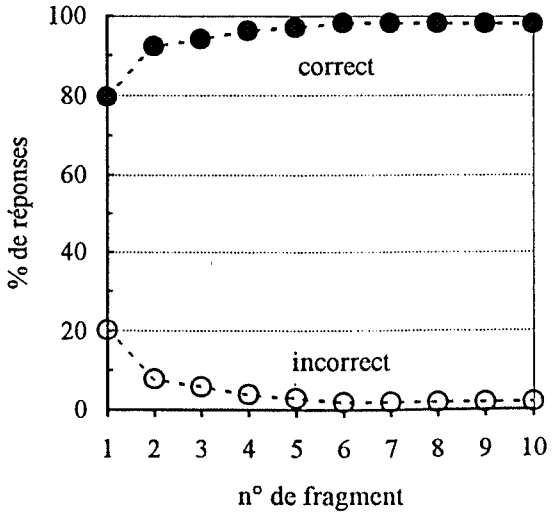


Figure 2 : Jugements corrects ou non selon le fragment (items contrôle).

Les résultats sont très nets : (a) une large majorité de jugements "dental" pour les items en /tl/ ou /dl/ ; (b) une baisse progressive des jugements "dental" au profit des "vélaire" à

partir du 4ème ou 5ème gate ; (c) pour les items contrôle, une précision des réponses croissant avec la durée présentée. La baisse (b) est d'environ 16% pour les 2 formats ; la corrélation entre pourcentage de réponses "dental" et numéro de gate, à partir du 4ème ou 5ème gate, est très significativement négative (format série : $r(54) = -.38$, $p = .004$; format parallèle : $r(46) = -.68$, $p < .0001$). Les fragments initiaux des clusters /tl/ ou /dl/ ont été jugés d'articulation dentale jusqu'à hauteur de 83% (série) ou 91% (parallèle) vers le 4ème gate, qui correspond à l'émergence de la détection de la liquide /l/ ; à partir de là, les sujets tendent à réviser leur jugement dans le sens vélaire. Ceci est observé aussi bien dans le format série où l'on pouvait craindre un effet de persévération (cf. Walley et al., 1995), que dans le format parallèle où cet effet est exclu.

Les résultats suggèrent donc que les clusters /tl/ et /dl/ ont une qualité dentale mais induisent une assimilation dental-vélaire de nature contextuelle. L'amplitude de l'effet en situation de gating est faible, sans doute parce que les fragments présentés sont trop courts pour que l'effet se développe pleinement. Nous avons donc cherché à le mesurer dans des situations où l'intégralité des items était présentée.

4. EXPÉRIENCES D'IDENTIFICATION

Deux tests d'identification avec choix forcé portant sur la consonne initiale ont été conduits. Dans le premier test, les sujets recevaient 2 blocs de stimuli, l'un avec initiale voisée, l'autre avec initiale non-voisée. Pour chaque bloc, le choix forcé était entre 3 lieux d'articulation [bdg] ou [ptk]. Les résultats, obtenus pour 24 sujets, sont résumés dans le Tableau 2. L'initiale des items en /tl/ et /dl/ était jugée vélaire dans plus de 55% des cas (comparé à 4% pour les items contrôles en /tr/ et /dr/), avec davantage de réponses "vélaire" pour les items en /tl/ qu'en /dl/. Le second test ne comportait plus l'alternative labiale, trouvée négligeable dans le premier test, mais permettait de tester les confusions de voisement. Un seul bloc de stimuli à initiale voisée ou non était présenté, le choix forcé étant [dtgk]. Les résultats confirment ceux du premier test : 50% de réponses 'g' pour les items en /dl/, 72% de réponses 'k' et 12% de réponses 'g' pour les items en /tl/ (16 sujets).

Tableau 2 : % d'identifications vélaire et labial pour les items de test et de contrôle.

jugement	*type de stimulus*			
	/tl/-	/dl/-	/tr/-	/dr/-
vélaire	63.5	47.9	7.3	1.0
labial	4.2	1.0	2.1	6.3

L'amplitude de l'effet d'assimilation est moindre que dans l'expérience préliminaire où les mêmes items étaient pourtant présentés de la même façon. La différence vient de la tâche. Dans les tâches d'identification, les sujets doivent se concentrer sur la consonne initiale au lieu de transcrire la totalité de chaque item ; d'autre part, il y a moins de chances que leurs réponses reflètent leurs connaissances orthographiques que pour la transcription libre qui favorise implicitement les graphèmes licites tels que "cl" (/kl/) ou "gl" en début de mot, plutôt que "tl" ou "dl".

La question se pose de savoir s'il y a encore assimilation perceptive dans la situation temps réel de détection de phonème . Les résultats de gating suggèrent plutôt que non, puisque les sujets perçoivent d'abord /d/ ou /t/ lorsqu'on leur présente le début de /dl/ ou /tl/. Il est aussi possible que les réponses de détection on-line nécessitent une intégration perceptive en unités plus larges que le phonème. Dans ce cas, les réponses de détection devraient déjà refléter l'assimilation "contextuelle" qui émerge dans les tâches de gating.

5. EXPÉRIENCE DE DÉTECTION

Nous avons utilisé pour cette expérience le paradigme de détection de phonème généralisée, où les phonèmes cibles peuvent apparaître dans n'importe quelle position (Frauenfelder et Segui, 1989). Les sujets devaient détecter des occlusives soit dentales soit vélaires dans les items critiques du type 'dlapto'. Si l'effet d'assimilation intervient déjà dans cette tâche on-line, on s'attend à des fausses alarmes pour les cibles vélaires et à des non-détections pour les cibles dentales. Les items en /tl/ et /dl/ étaient complétés par des items contrôle en /kl/ et /gl/ (e.g., 'glapto' pour 'dlapto') associés aux mêmes cibles dentales ou vélaires, pour permettre de comparer fausses

alarmes et détections correctes pour les cibles /g/ et /k/. Les stimuli présentés étaient groupés par cible, avec environ un stimulus sur 3 portant la cible. Par exemple, le bloc pour la cible /k/ contenait 8 stimuli avec /k/ en initiale (items en /kl/ et en /kr/), et 16 stimuli avec /k/ en position médiale ou finale ; ce bloc contenait aussi les 4 items en /tl/, destinés à induire des fausses alarmes, et 48 items de remplissage. Un total de 48 sujets ont participé à l'expérience. Les pourcentages de détection correcte (hits) et de fausses alarmes (f.a.) sont récapitulés dans le Tableau 3.

Tableau 3 : Pourcentages de détection en fonction de la cible et du type d'item.

Type d'item	*cible dentale*	*cible vélaire*
/dl/- ou /tl/-	33.8%	79.7%
/gl/- ou /kl/-	5.2%	83.8%

Les cibles /g/ et /k/ étaient donc à peu près aussi souvent détectés dans les items en /dl/ et /tl/ que dans ceux en /gl/ et /kl/ (environ 80% du temps). Ce résultat montre que l'assimilation se produit aussi dans une tâche on-line, avec une amplitude plus forte encore que dans les tâches d'identification off-line.

Mais y a-t-il une difficulté particulière à percevoir par exemple /k/ dans /tl/ plutôt que dans /kl/ ? Cette difficulté n'apparaît pas dans les pourcentages de détection (79.7% *vs.* 83.8%) mais dans les temps de réponse, significativement plus longs pour les fausses alarmes de 34 ms en moyenne (voir Figure 3).

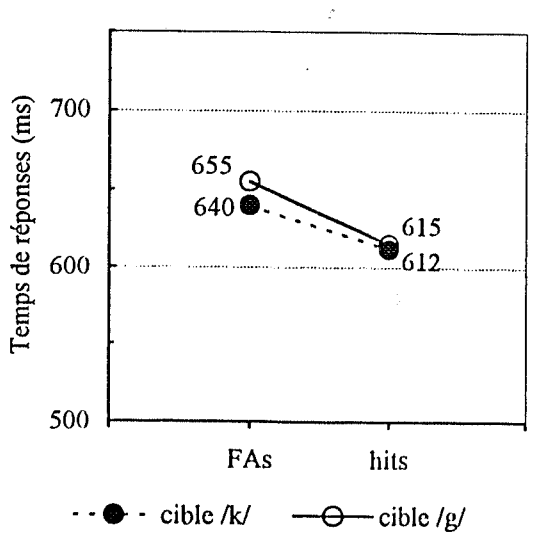


Figure 3 . Temps de détection de /k/ et /g/ dans /tl/ et /dl/ (f.a.) *vs.* /kl/ et /gl/ (hits).

6. CONCLUSION

Cette série d'expériences montre que les francophones ont une forte tendance à entendre /kl/ et /gl/ au lieu de /tl/ et /dl/ en position initiale de mot. Ce phénomène est cohérent avec la notion d'un filtre phonologique incluant les contraintes phonotactiques de la langue et montre que les transgressions /tl/ et /dl/ ne sont pas "transparentes" : /tl/ et /dl/ sont assimilés à des clusters légaux. L'assimilation apparaît dès que la composante liquide des clusters /tl/ ou /dl/ est perçue. C'est dans ce sens qu'elle peut être qualifiée de contextuelle.

Ce type d'assimilation s'accompagne d'un coût perceptif : les temps de détection des vélaires finalement perçues sont plus longs dans les clusters illégaux /tl/ et /dl/ (f.a.) que dans les clusters légaux /kl/ et /gl/ (hit).

7. BIBLIOGRAPHIE

Dupoux, E., Kakehi, K., Hirose, Y., Pallier, C., Fitneva, S., & Mehler, J. (sous presse). Epenthetic vowels in Japanese: a perceptual illusion? *Journal of Memory and Language*.

Frauenfelder, U., & Segui, J. (1989). Phoneme monitoring and lexical processing: Evidence for associative context effects. *Memory and Cognition*, **17**.

Grosjean, F. (1980). Spoken word-recognition processes and the gating paradigm. *Perception and Psychophysics*, **28**, 267-283.

Massaro, D., & Cohen, M. (1983). Phonological context in speech perception. *Perception and Psychophysics*, **34**, 338-348.

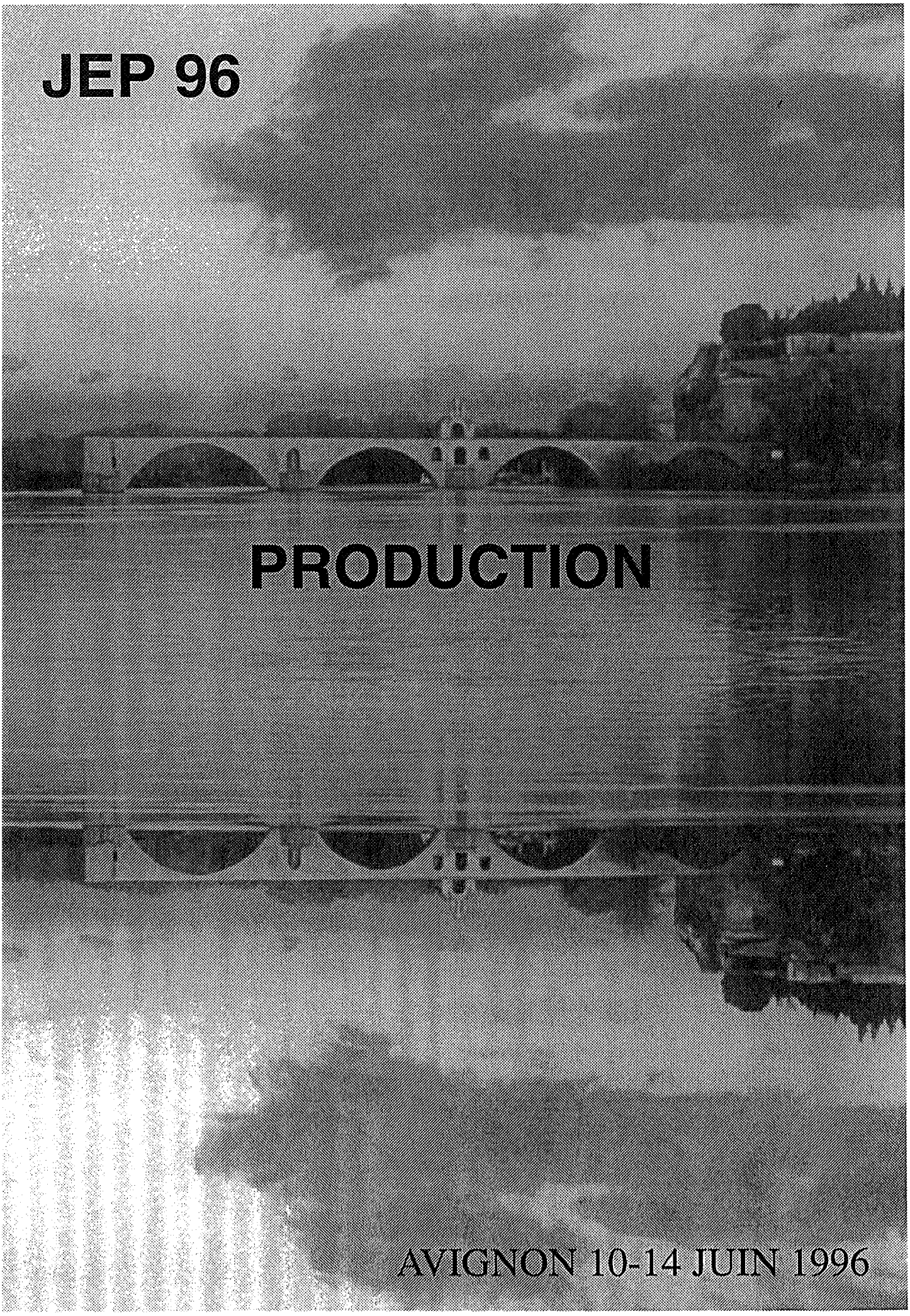
Polivanov, E. (1931). La perception des sons d'une langue étrangère. *Travaux du Cercle Linguistique de Prague*, **4** (reproduit en partie dans *Change*, 1969, **3**, 111-114).

Robel, L. (1969). Polivanov et le concept de surdité phonologique. *Change*, **3**, 115-119.

Walley, A.C., Michela, V.L., & Wood, D.R. (1995). The gating paradigm: Effects of presentation format on spoken word recognition by children and adults. *Perception and Psychophysics*, **57**, 343-351

Warren, R.M. (1970). Perceptual restoration of missing speech sounds. *Science*, **167**, 393-395.

JEP 96



PRODUCTION

AVIGNON 10-14 JUIN 1996

AMULET : UN SYTEME D'ANNOTATION AUTOMATIQUE DE DONNEES MULTISENSORIELLES

Nathalie Parlangeau

Institut de Recherche en Informatique de Toulouse 118, Route de Narbonne, 31062 Toulouse Cédex

Tel : 61 55 60 55 - Fax : 61 55 62 58 - e-mail : parlange@irit.fr

ABSTRACT

Speech production is a complex process relying on coordinated gestures, but the acoustic signal does not depict its underlaying organization. Accepting that articulatory gestures are directly recognized through the coarticulation process, our proposal is to investigate the correlations between acoustic and articulatory informations and to assess gestural phonetic theory. We present here the framework for this investigation, the automatic labelling of the multi-sensor speech database ACCOR in terms of articulatory events. Automatic procedures of annotation are based on statistic methods of segmentation and are mainly issued from signal processing domain.

1. INTRODUCTION

Il est notoire que le signal acoustique de parole est extrêmement variable ; cette caractéristique est certainement l'une des raisons pour lesquelles les systèmes de Reconnaissance Automatique de la Parole (RAP) atteignent actuellement des limites quant à leur performance. Une des causes de cette variabilité est due au phénomène de coarticulation qui résulte d'une combinaison de contraintes physiologiques et linguistiques liées à la structure phonologique d'une langue donnée (Vaissière, 1986). La production de la parole est un processus complexe qui repose sur des événements articulatoires coordonnés, mais le signal produit ne reflète pas cette structure sous-jascente. D'un point de vue théorique, plusieurs chercheurs voient dans l'articulation, un niveau de représentation intermédiaire qui permet de lier production et perception. Une théorie, à savoir la phonologie gesturale, fait l'hypothèse que les événements articulatoires sont directement reconnus au travers du processus de coarticulation. Afin de valider cette approche, il est nécessaire d'étudier des signaux articulatoires et acoustiques, de déterminer précisément les événements articulatoires de base, d'en extraire les corrélations pour finalement en déduire l'organisation temporelle. Si cette théorie s'avère exacte, elle doit permettre de préciser ultérieurement un niveau de représentation intermédiaire dans des modèles de type statistique destinés à la RAP.

Au cours des travaux présentés dans cette communication, nous avons repris les annotations de type articulatoire proposées par les phonéticiens et nous avons cherché à réaliser cet étiquetage en termes d'événements articulatoires élémentaires,et ce de manière automatique. Les signaux étudiés sont issus de la base de données multisensorielles ACCOR (Articulatory Correlations of CoaRticulation patterns) (Marchal, 1993), cette base regroupe des données acoustiques, aérodynamiques et articulatoires : signal acoustique, signal laryngographique, signaux du débit d'air nasal et oral ainsi que l'ElectroPalatoGramme.

L'ensemble des procédures automatiques du système AMULET (Automatic MUltisensor Labelling and Event Tracking) sont basées sur des méthodes issues du domaine du traitement du signal appliquées dans un cadre méthodologique.

2. CADRE METHODOLOGIQUE

L'étiquetage de la base de données est basée sur deux principes fondamentaux :

- **la non-linéarité** ou indépendance horizontale,

- **l'indépendance orthogonale des annotations** sur les différents signaux.

La non-linéarité : la production de la parole est un processus complexe qui repose sur des gestes articulatoires coordonnés. Par essence, c'est un mécanisme non-linéaire : le signal de parole ne peut pas être considéré comme une simple concaténation de segments et le phénomène de coarticulation ne se réduit pas à un processus d'accomodation, les gestes articulatoires se chevauchent et les événements résultants peuvent être associés à plusieurs unités phonétiques , comme le met en évidence l'étude de l'organisation des contacts linguo-palataux (Marchal, 1990).

L'indépendance orthogonale : le but final de ces travaux est l'étude de la corrélation entre les événements articulatoires repérés sur les différents signaux articulatoires et le signal acoustique. Il est impératif, à moins de risquer de biaiser cette étude, d'annoter chaque signal de façon indépendante afin de permettre ensuite une interprétation systématique des corrélations

entre les différents signaux et de définir une réelle coordination entre les événements.

L'automatisation des procédures d'annotation n'a d'intéret par rapport à l'annotation manuelle, que si les événements sont trouvés de manière systématique et sans erreur. Il convient donc d'ajouter un troisième principe : **la robustesse des méthodes** dans le sens où les détections doivent être indépendantes du locuteur et consistentes.

Les méthodes d'annotation que nous utilisons suivent toutes le même schéma : dans un premier temps, les discontinuités du signal sont détectées, seules certaines sont retenues et ensuite interprétées comme des indications de événements articulatoires. Les événements articulatoires sont marqués dans le domaine temporel selon des critères relevant de connaissances phonétiques.

3. ANNOTATION DU SIGNAL ACOUSTIQUE

Les événements recherchés par les experts phonéticiens sur le signal acoustique sont liés au voisement et aux fermetures du conduit vocal. Le repérage de ces événements se ramène à la détection des événements :

- VOW et VTW, qui marquent le début et la fin de voisement,

- SRW, synonyme du relâchement du conduit vocal lors de la prononciation des occlusives sourdes.

3.1. Les méthodes de détection

3.1.1. Méthode automatique de segmentation

Dans un premier temps, les discontinuités acoustiques sont détectées à partir dune méthode de segmentation automatique robuste, la méthode de divergence Forward-Backward (André-Obrecht, 1988). Le signal est supposé être décrit par une séquence d'unités stationnaires caractérisées par un modèle autorégressif gaussien. La méthode consiste à détecter les changements des paramètres du modèle. Le test de divergence est basé sur la mesure d'une distance entre deux modèles autorégressifs. Une rupture est détectée lorsque cette distance dépasse un certain seuil. Deux procédures de détection, basées sur cette méthode, sont mises en oeuvre en parallèle, une sur le signal filtré (hautes fréquences) et une sur le signal non filtré. Afin d'éviter les omissions, une procédure de détection est mise en oeuvre dans le sens rétrograde. Les paramètres (ordre des modèles, seuils) sont indépendants du locuteur, et n'ont pas nécessité d'apprentissage sur cette base de données.

La segmentation obtenue est infra-phonémique. Afin de ne garder que les frontières de segment significatives des événements articulatoires, nous avons développé un test de voisement, et un test de détection du relâchement des occlusives.

3.1.2. Le test de voisement

Le test de voisement utilisé est basé sur trois paramètres : l'énergie du signal, l'autocorrélation du signal, et le premier coefficient de réflection. Le résultat de ce test est corrigé par le Taux de Passages à Zéro.

Nous caractérisons ainsi tous les segments obtenus comme voisés ou non voisés. Les segments de même nature sont regroupés et les frontières extrêmes donnent les évènements VOW et VTW.

3.1.3 Le test de détection des occlusives.

Bien que la segmentation automatique détecte généralement le relâchement de l'occlusion, nous avons développé un test complémentaire basé sur une analyse centiseconde du signal afin de rendre la détection plus robuste. Le test utilise deux paramètres, l'énergie formantique du signal, Efn, énergie du signal dans la zone fréquentielle [500Hz-3000Hz] et l'énergie résiduelle du signal, sous l'hypothèse autorégressive. L'énergie formantique permet de déterminer les zones d'occlusion potentielles du conduit vocal. Un modèle autorégressif gaussien est identifié sur chaque zone, l'énergie résiduelle est calculée au niveau de la frontière droite sur des fenêtres glissantes de 2ms : un franchissement de seuil par cette énergie est équivalent à une ouverture brutale du conduit vocal. Les seuils utilisés sont adaptatifs, ils sont évalués à partir de l'écart-type des paramètres sur les zones de recherche. Le test de voisement est ensuite appliqué de part et d'autre de la détection afin de valider le relâchement d'une occlusive SRW.

3.2. Résultats et discussion

Le corpus d'évaluation pour le français est actuellement composé de cinq répétitions de deux phrases par deux locuteurs : " La cousine de Vichy épousa un hippie à Toulouse " et " C'est maintenant que la smala les acclame ". Le corpus anglais est composé d'une répétition de trois phrases par trois locuteurs.

L'évaluation consiste à dénombrer les omissions et les insertions, et, lorsqu'il y a détection justifiée, à mesurer le délai entre l'étiquetage manuel et l'étiquetage automatique. Les résultats de l'évaluation ne sont détaillés que pour le corpus français. Les résultats obtenus sur le corpus anglais confirment les résultats obtenus sur le corpus français.

Tableau 1 : étiquetage manuel vs étiquetage automatique. Délais en ms.

	< 10	10 < 20	> 20	Inséré	Omis
VOW	111/129	4/129	12/129	5	2/129
VTW	106/129	11/129	10/129	6	2/129
SRW	51/73	5/73	1/73	18	16/73

Les délais supérieurs à 20 ms pour la fin de voisement VTW (tableau 1), sont souvent expliqués par une onde sinusoidale persistante sur le signal. Pour la détection de SRW, notre méthode détecte tous les relâchements des phonèmes /t/ et /k/, ainsi que le relâchement de l'occlusive labiale /p/ dans un contexte hautes fréquences.

4. ANNOTATION DU SIGNAL LARYNGOGRAPHIQUE

Le laryngographe a pour but de mettre en évidence le mouvement des cordes vocales et du larynx. Les événements recherchés sur ce signal sont donc :

- VOX et VTX qui sont respectivement le début et la fin de vibration des cordes vocales,

- PUX qui est un pic sur un segment non voisé du signal, révélateur d'un mouvement brusque du larynx.

4.1. Les méthodes de détection

Dans un premier temps, les principales discontinuités sont détectées grâce à une version simplifiée de la méthode Forward-Backward. Un test de voisement adaptatif permet ensuite de caractériser les frontières obtenues en terme de début et fin de voisement, et les variations du gradient nous permettent de détecter le pic sur les segments non voisés.

4.1.1. Le test de voisement adaptatif

Le signal laryngographique n'est pas de moyenne nulle, celle ci est donc calculée sur une fenêtre glissante afin de la retrancher au signal et appliquer ensuite la méthode Forward-Backward. Sur chaque segment, deux niveaux sont définis de part et dautre de la moyenne du signal et nous mesurons le nombre de passages par ces niveaux. Nous obtenons ainsi un nombre important de passages à niveaux pour les segments voisés et faible pour les segments non voisés. Ce test s'est révélé très robuste. Les frontières des segments voisés sont interprétées en terme de début et fin de voisement.

4.1.2. Le test de détection du PUX

Sur les segments non voisés, une recherche systématique des changements de pente nous permet simplement de mettre en évidence les pics significatifs que nous étiquetons PUX..

4.2. Résultats et discussion

Nous obtenons d'excellents résultats pour les débuts et fins de voisement (tableau 2). Pour le PUX, 9 événements ne sont pas détectés, et l'on observe quelques insertions. Ces résultats sont dûs à un manque de précision des critères d'annotation manuelle pour un tel événement.

Tableau 2 : étiquetage manuel vs étiquetage automatique. Délais en ms.

	< 10	10 <20	>·20	Inséré	Omis
VOX	115/125	6/125	2/125	1	2/125
VTX	118/125	3/125	2/125	6	2/125
PUX	56/66	1/66		27	9/66

5. ANNOTATION DES SIGNAUX AÉRODYNAMIQUES : LE SIGNAL NASAL ET LE SIGNAL ORAL

Les événements recherchés sur le signal nasal sont :

- BFN et DFN qui sont respectivement le début et la fin d'une activité nasale,

- MFN le maximum de débit d'air. Cette étiquette est souvent couplée à la fin de l'activité nasale, auquel cas on a l'étiquette MFDFN.

Les événements recherchés sur le signal oral sont :

- BFO et DFO qui sont respectivement le début et la fin d'une activité orale,

- MFO et mFO sont respectivement le maximum et le minimum d'activité orale,

- SCO et SRO sont respectivement l'occlusion et le relâchement d'une occlusive.

5.1. Les méthodes de détection

Les événements recherchés sur ces signaux correspondent à des variations de pente que nous obtenons par interpolation linéaire. La segmentation des signaux correspond à des détections de débuts et fins de descente. Un ensemble de règles spécifiques nous permet d'interpréter ces changements en terme d'événements articulatoires.

La recherche des événements SCO et SRO est particulière : SCO est un événement lié à une forte diminution du débit d'air, pendant un temps plus ou moins long. Ce débit doit descendre sous un seuil atteint quasiment exclusivement lors de la réalisation du relâchement des occlusives.

5.2. Résultats et discussion

L'annotation du signal nasal est difficile, l'étiquetage manuel de référence étant souvent trop fin (tableau 3). Ceci explique la plupart des omissions. De plus, certains critères peuvent varier suivant les experts, ce qui explique la plupart des insertions.

Tableau 3 : étiquetage manuel vs étiquetage automatique. Délais en ms.

	< 10	10<20	> 20	Inséré	Omis
BFN	83/119	6/119	19/119	7	11/119
MFN	99/117	5/117	5/117	10	8/117
DFN	92/118	2/118	11/118	4	13/118

Pour le signal oral, dans un délai inférieur à 20 m, les résultats sont bons (tableau4). On peut remarquer une déficience de l'annotation automatique du SCO, résultat de critères trop subjectifs fournis par les phonéticiens

Tableau 4 : étiquetage manuel vs étiquetage automatique. Délais en ms.

	< 10	10<20	>20	Inséré	Omis
BFO	49/93	16/93	9/93	5	19/93
DFO	64/101	23/101	4/101	6	10/101
MFO	67/83	12/83	1/83	40	3/83
mFO	19/53	16/53	3/53	2	21/53
SCO	17/32	4/32	2/32	3	9/32
SRO	27/41	3/41	5/41	1	6/41

6. ANNOTATION DU SIGNAL ELECTROPALATOGRAPHIQUE

L'E.P.G. est une succession d'images de 8*8 points, représentant les contacts de la langue avec le palais. Les phonéticiens recherchent sur ce signal des événements liés à la réalisation d'occlusions :

- ACE l'approche d'occlusion,
- SCE le moment d'occlusion,
- MCE le maximum d'occlusion,
- SRE le relâchement d'occlusion.

6.1. Les méthodes de détection

L'étiquetage de l'E.P.G. est un processus de recherche dynamique à partir de la localisation des moments précis d'occlusion. Pour détecter les aires d'occlusion, nous utilisons deux masques correspondant aux deux endroits possibles de réalisation des occlusions sur le palais (palatale ou vélaire). Les frontières d'une zone d'occlusion nous donnent les moments exacts d'occlusion et de relâchement SCE et SRE. L'approche ACE est recherchée avant le moment de fermeture, selon l'endroit de réalisation de l'occlusion sur le palais : elle correspond à un nombre de contacts suffisant vers le centre. Le maximum MCE est la première image pour laquelle le nombre de contacts est maximum dans la zone d'occlusion.

6.2. Résultats et discussion

La détection des événements SCE et SRE est très robuste et très précise en temps et lieu d'articulation (tableau 5). La détection de l'approche d'occlusion est dépendante du contexte la précédant, ce qui implique des détections avec des délais supérieurs à 20 ms. L'étiquetage manuel de l'événement MCE peut répondre à différentes stratégies ce qui explique les délais supérieurs à 10 ms.

Tableau 5 : étiquetage manuel vs étiquetage automatique. Délais en ms.

	< 10	10 < 20	> 20
ACE	78/124	8/124	38/124
SCE	124/124		
MCE	110/118	4/118	4/118
SRE	124/124		

7. CONCLUSION

Nous avons défini un système automatique d'annotation d'une base de données multi-sensorielles, et les résultats obtenus sont bons : environ 80 % de bonnes détections pour les quatre premiers signaux, et environ 90% de bonnes détection pour l'E.P.G.. Quelques écarts d'annotation sont dûs à notre système automatique, et d'autres sont dûs aux variations des critères de l'annotation manuelle.

La mise au point des procédures automatiques de segmentation a permis à la fois de confirmer et de préciser les critères de segmentation manuels, ainsi que de revoir la signification de certains événements au travers d'un échange avec les experts phonéticiens.

Ce travail est interessant dans la mesure où il permet d'assurer l'application stricte de critères précis et ce de façon robuste. Il permettra d'annoter un grand nombre de données dans un laps de temps raisonnable. A partir de ces annotations, nous étudierons la coordination temporelle des événements acoustiques et articulatoires mis en oeuvre dans la production des plosives sourdes. Cela nous permettra d'étudier une alternative à de précédents modèles articulatoires pour la reconnaissance automatique de la parole (Erler, 1992).

8. REFERENCES

Vaissière J (1986), *Speech recognition : a tutorial* , ed. F. Fallside and W. A. Woods, Prentice Hall International, pp 191-236.

Marchal A., Hardcastle W.J. (1993), ACCOR : Instrumentation and database for cross-langage study of coarticulation , Langage and Speech, 2-3,pp 137-153,.

Marchal A., N'Guyan-Trong N. (1990), " Non Linearity and phonetic Segmentation", J. Acoust. Soc. Am., Suppl.1, Vol87, pp79-82.

André-Obrecht R. (1988), " A new approach for the automatic segmentation of continuous speech signals", *IEEE Trans on ASSP.*

Erler K, Deng L. (1992), " HMM representation of quantized articulatory features for recognition of highly confusible words", *ICASSP 92*, Vol 1, pp 545-548.

CONSÉQUENCES ACOUSTIQUES DU PASSAGE DE LA COUPE SAGITTALE À LA FONCTION D'AIRE

Véronique Lecuit[1] et Alain Soquet[2]

Laboratoire de Phonétique Expérimentale – Université Libre de Bruxelles
50 av. F.-D. Roosevelt, CP110, B-1050 Bruxelles– Belgique
Tél.: (32 2) 650 20 18 – Fax: (32 2) 650 20 07 – e-mail vlecuit@ulb.ac.be
[1]Bourse U. L. B., [2]Projet de la Communauté Française de Belgique, ARC 93/98-168.

ABSTRACT

The generation of area functions from measurements of the sagittal section is an important step in the study of the relation between vocal tract geometry and speech acoustics. Many authors have proposed transformations performing this particular task. In this paper, we carried out a comparative study of four of these transformations. We used two sets of sagittal cuts, one obtained by means of X-rays and the other generated with Maeda's model. The outcome of this study is that the choice of a particular transformation has important influence both on vocal tract area function and on acoustic cues.

1. INTRODUCTION

Il existe plusieurs modèles qui tentent de décrire les mécanismes de la production de la parole. Entre autres, des modèles articulatoires décrivent la forme du conduit vocal en termes d'un petit nombre de paramètres tandis que des modèles acoustiques décrivent la propagation des ondes sonores dans le conduit vocal.

La validité de ces modèles est vérifiée grâce aux observations directes du conduit effectuées au moyen de diverses techniques. Les données articulatoires les plus courantes sont des coupes sagittales du conduit vocal obtenues par rayons X, ou par résonance magnétique nucléaire. Dans le cas des rayons X, les données acoustiques peuvent être obtenues par enregistrement du signal de parole effectué lors de la prise des mesures articulatoires. A partir du signal de parole, il est possible d'extraire des indices acoustiques (par exemple les trois premiers formants). Pour pouvoir étudier les relations entre les indices articulatoires et acoustiques ainsi obtenus, la coupe sagittale, représentation à deux dimensions du conduit vocal, doit être transformée en une représentation à trois dimensions, la fonction d'aire. La fonction d'aire représente la variation de l'aire de la section transversale du conduit vocal de la glotte aux lèvres. Pour ce faire, plusieurs auteurs ont proposé des transformations mathématiques (Perrier et al. [8], Fant [2], Sorokin [9], et Sundberg et al. [10]). Ces transformations étant fréquement utilisées, il est intéressant de les étudier systématiquement.

2. TRANSFORMATIONS

Nous avons choisi de comparer les quatre transformations mentionnées ci-dessus. Les auteurs proposent des transformations permettant de passer du diamètre de la section relevé sur la coupe à l'aire de la section correspondante. Le conduit vocal est subdivisé en différentes zones (voir figure 1). Sundberg, Perrier et Fant proposent une formule du type

$$A(x) = \alpha(x)\, d(x)^{\beta(x)}$$

où A est l'aire calculée, d le diamètre sagittal et x la distance à la glotte. Les coefficients a et b sont ajustés par les auteurs suivant la zone du conduit vocal où l'on se situe. Sorokin propose une transformation d'un tout autre type: il utilise des formules ad hoc suivant les régions du conduit vocal. Certains auteurs ne traitent pas la zone labiale; pour ceux-ci, nous avons utilisé la transformation spécifique à cette zone proposée par Lindblom et Sundberg [6].

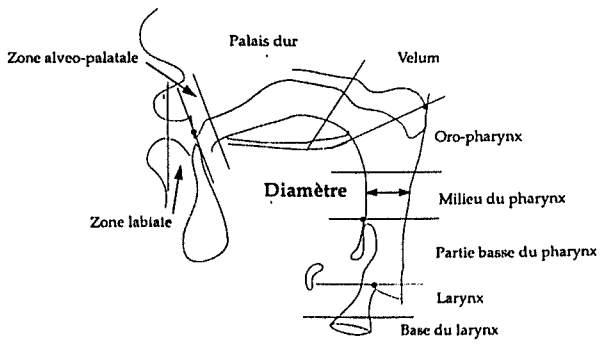


Figure 1: Représentation des différentes zones du conduit vocal.

3. DIGITALISATION

Les coupes sagittales utilisées sont les tracés obtenus d'après des clichés rayons X publiés par Bothorel et al. [1]. Les tracés de dix voyelles du français prononcées par un locuteur masculin (A. D.) et un locuteur féminin (P. B.) ont été utilisés. Etant donné que toutes les transformations calculent la fonction d'aire à partir du diamètre sagittal, nous avons développé un logiciel permettant la digitalisation des coupes et l'extraction automatique des diamètres.

Un des problèmes majeurs consiste à définir un système de coordonnées qui épouse au mieux la forme du conduit vocal; nous avons adopté un système de coordonnées de type semi-polaire comme proposé par Heinz et Stevens [3]. La figure 2 représente un tracé digitalisé sur lequel se superpose le système de coordonnées, différents points de repère permettent de définir les zones du conduit vocal. Le logiciel permet de définir différents paramètres, tels l'angle θ entre les deux sections droites, l'écart entre les lignes du système, la taille du système.

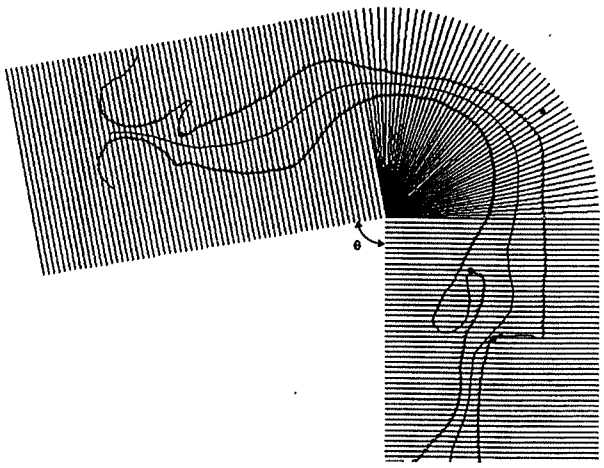


Figure 2: Tracé digitalisé sur lequel se superpose le système de coordonnées.

Les diamètres sagittaux sont calculés à partir des intersections entre la coupe et le système de coordonnées. Le profil sagittal s'obtient en disposant les diamètres sagittaux de la glotte aux lèvres.

4. FONCTIONS D'AIRE

Pour un profil sagittal donné, les quatre transformations permettent d'obtenir les fonctions d'aire correspondantes.

La figure 3 (a) décrit le profil sagittal de la voyelle [a] prononcée par le locuteur masculin. La figure 3 (b) décrit les quatre fonctions d'aire obtenues à l'aide des différentes transformations pour ce même profil sagittal.

L'étude de ces données permet de distinguer quelques différences entre les transformations. La transformation proposée par Perrier et al. [8] tend à surestimer les grandes sections tandis que celle proposée par Sorokin [9] sous-estime toutes les sections. Pour chaque voyelle et chaque locuteur, nous avons obtenu des résultats similaires. On peut observer que toutes les transformations permettent de localiser correctement les constrictions, exceptée celle proposée par Sorokin: en effet, la fonction d'aire étant peu contrastée, il est parfois difficile de les localiser. Le degré de constriction est souvent sous-estimé par Fant [2] et Sundberg et al. [10] tandis que Perrier et al. [8] surestime son importance.

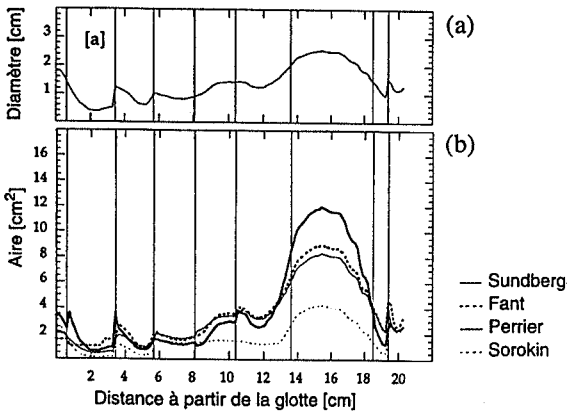


Figure 3: (a) Profil sagittal et (b) fonctions d'aire obtenues au moyen des quatre transformations pour la voyelle [a], prononcée par le locuteur masculin.

Il est difficile de tirer des conclusions définitives à partir de ces observations. En effet, nous ne pouvons pas comparer ces fonctions d'aire à la fonction d'aire "réelle".

Cependant, nous avons à notre disposition une estimation des formants correspondant à cette fonction d'aire réelle.

5. FORMANTS

A partir de la fonction d'aire, nous avons calculé les formants afin de pouvoir les comparer aux formants de référence extraits

de l'analyse LPC publiée par Bothorel et al. [1].

La figure 4 et la figure 5 présentent les formants obtenus par la méthode variationelle (Jospa [4]) pour chaque voyelle prononcée respectivement par le locuteur masculin et le locuteur féminin.

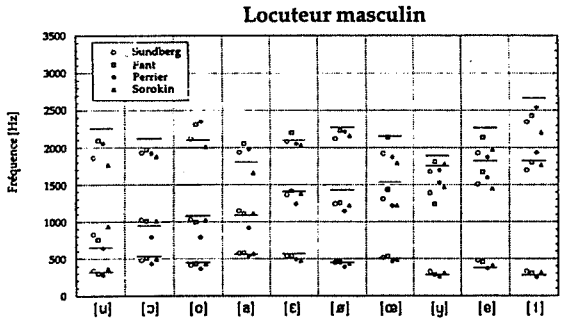


Figure 4: Représentation des valeurs des trois formants pour les 10 voyelles prononcées par le locuteur masculin.

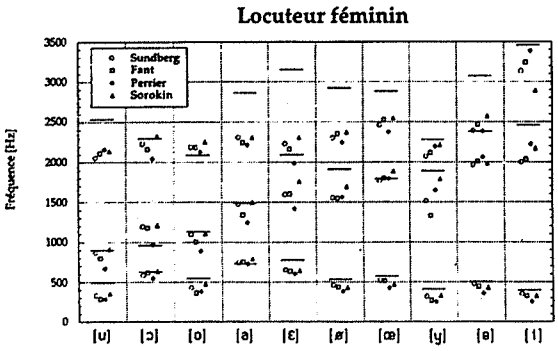


Figure 5: Représentation des valeurs des trois formants pour les 10 voyelles prononcées par le locuteur féminin.

La fréquence, représentée en ordonnée, permet de localiser les trois premiers formants. Les segments horizontaux représentent les formants de référence, les symboles représentent les formants obtenus après application des différentes transformations.

On remarque que pour F1, les fréquences calculées sont très proches des fréquences de référence. Cependant, les fréquences des deuxième et troisième formants sont rarement en accord avec la référence. Néanmoins, il convient de noter l'importance du positionnement du système de coordonnées semi-polaires. En effet, de petites variations sur la position du système de coordonnées peuvent influencer sensiblement le profil sagittal et donc également la fonction d'aire et les formants.

Nous avons donc pensé qu'il serait intéressant de comparer les formants obtenus par les différentes transformations entre eux, et non plus par rapport aux formants de référence.

De manière à augmenter le nombre de coupes nous avons généré 15 000 coupes sagittales au moyen du modèle de Maeda [7]. Les fonctions d'aire et les formants ont été calculés à l'aide des quatre transformations. Nous avons ensuite comparé les transformations deux à deux pour chaque formant et pour les 15 000 coupes. Parmi l'ensemble des résultats (Lecuit [5]), nous avons sélectionné quelques tendances caractéristiques.

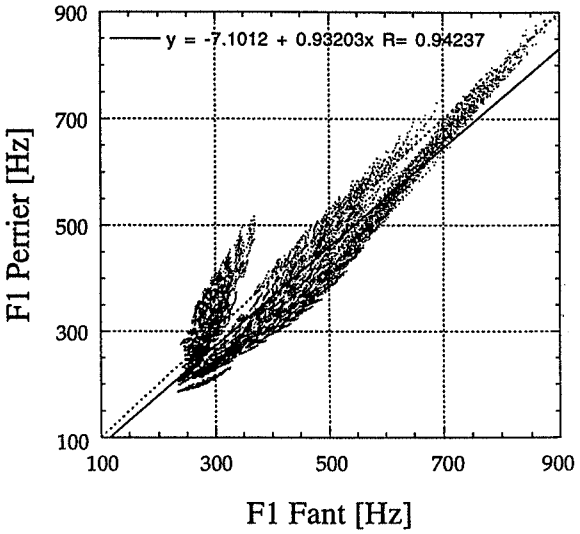


Figure 6: Représentation du F1 obtenu par Fant versus le F1 obtenu par Perrier pour les 15 000 coupes (la droite de régression ainsi que la droite de pente unité et passant par l'origine sont représentées).

Nous avons observé que Fant [2] conduit à sous-estimer les F1 compris entre 250 et 350 Hz. Ce comportement s'observe non seulement par rapport à Perrier et al. [8] comme illustré à la figure 6, mais également par rapport aux deux autres transformations.

Toutefois on n'observe pas toujours des comportements aussi spectaculaires: par exemple à la figure 7 on peut observer une relation quasi-linéaire entre les F2 de Sundberg et al. [10] et Perrier et al. [8].

A la figure 8 on peut noter la difficulté qu'a Sorokin [9] d'obtenir des F3 supérieurs à 3000 Hz.

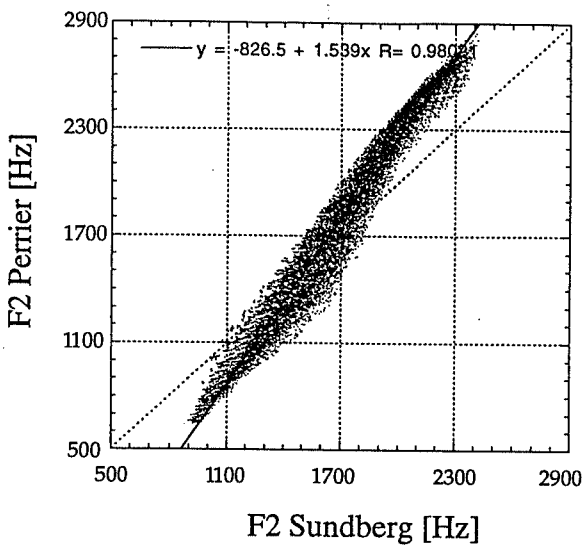


Figure 7: Représentation du F2 obtenu par Sundberg versus le F2 obtenu par Perrier pour les 15 000 coupes (la droite de régression ainsi que la droite de pente unité et passant par l'origine sont représentées).

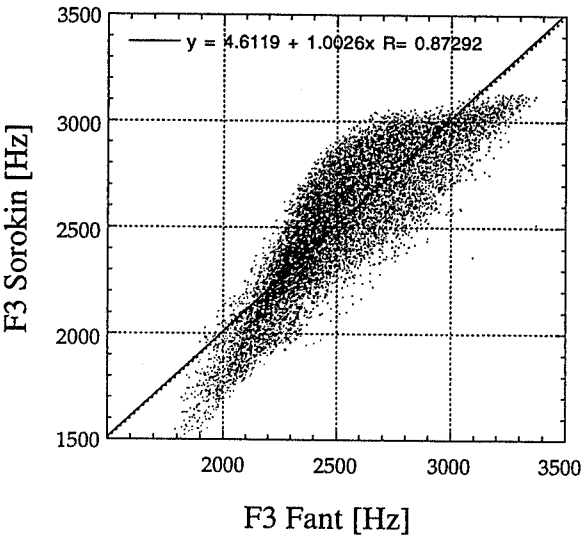


Figure 8: Représentation du F3 obtenu par Fant versus le F3 obtenu par Sorokin pour les 15 000 coupes (la droite de régression ainsi que la droite de pente unité et passant par l'origine sont représentées).

6. CONCLUSION

Au terme de cette étude on constate que les différentes transformations ont une influence sur les résultats acoustiques. Quand on remonte à l'origine de ces transformations on remarque qu'elles ont été élaborées à partir de données réelles assez restreintes. Les paramètres ont parfois été ajustés sur un locuteur unique. On ne peut donc les généraliser sans risques.

Cependant, ces transformations sont couramment utilisées, et à l'heure actuelle on ne peut s'en passer. Il faut donc être conscient de leur influence lorsque l'on utilise ces transformations dans d'autres types d'études.

7. BIBLIOGRAPHIE

[1] A. Bothorel, P. Simon, F. Wioland, and J. P. Zerling, Cinéradiographie des voyelles et consonnes du français, Travaux de l'Institut de Phonétique de Strasbourg, Strasbourg, 1986.

[2] G. Fant, "Vocal tract area functions of Swedish vowels and a new three-parameter model", Proc. 1992 Internat. Conf. on Spoken Language Processing, Banff, Canada, vol. 1, Paper Fr.fAM.3.1, pp. 807-810, 1992.

[3] J. M. Heinz et K. N. Stevens, "On the relations between lateral cineradiographs area functions, and acoustic spectra of speech", 5th. Int. Congr. Acoust., Paper A44, Liège, 1965.

[4] P. Jospa, "Formulation variationelle du lien acoustico-articulatoire", Actes des 20èmes Journées d'Etude sur la Parole, Trégastel, pp. 113-118, 1994.

[5] V. Lecuit, "Sagittal cut to area function transformations: a comparative study," Mémoire, Faculté des Sciences et Institut des Langues Vivantes et de Phonétique, Université Libre de Bruxelles, 1995.

[6] B. Lindblom et J. Sundberg, "Acoustical consequences of lip, tongue, jaw and larynx movement", J.A.S.A., vol. 50, pp. 1166-1179, 1971.

[7] S. Maeda, "Un modèle articulatoire de la langue avec des composantes linéaires", Actes des 10èmes Journées d'Etude sur la Parole, Grenoble, pp. 154-162, 1979.

[8] P. Perrier, L. J. Boë et R. Sock, "Vocal tract area function estimation from midsagittal dimensions with CT scans and a vocal tract cast: Modeling the transition with two sets of coefficients", J. Speech Hearing Res., vol. 35, pp. 53-67, 1992.

[9] V. N. Sorokin, "Determination of vocal tract shape for vowels", Speech Communication, vol. 11, pp. 71-85, 1992.

[10] J. Sundberg, C. Johansson, H. Wilbrand, and C. Ytterbergh, "From sagittal distance to area. A study of transverse, vocal tract cross-sectional area", Phonetica, vol. 44, pp. 76-90, 1987.

CONTRÔLE DE LA LANGUE EN PAROLE : QUELQUES PROPOSITIONS TESTÉES SUR UNE MODÉLISATION BIOMÉCANIQUE

Yohan PAYAN & Pascal PERRIER

Institut de la Communication Parlée - URA CNRS 368 - INPG & Université Stendhal
46 avenue Félix Viallet, 38031 Grenoble Cedex 1, France
Tél.: 76 57 47 14 - Fax : 76 57 47 10 - e-mail : payan@icp.grenet.fr

ABSTRACT

This paper aims to focus on the use of Feldman's Equilibrium Point Hypothesis for the control of tongue movements in speech. In this perspective, a Finite Element model of mid-sagittal plan tongue motion based on this theory, has been developed. We assess here the ability of our model to reproduce kinematics properties of a speaker tongue movement in a [i-a] transition. The articulatory and acoustic simulations of the transition are then compared to the recorded ones.

Key words : speech production, articulatory and biomechanical modeling, motor control.

1. INTRODUCTION

Comme le soulignaient Boë, Maeda et Perrier [1] lors des dernières Journées d'Étude sur la Parole, la modélisation articulatoire a vu un demi siècle d'évolution, apportant à de nombreux chercheurs du domaine de la production de la parole, une meilleure connaissance des processus physiologiques à l'origine du signal de parole ainsi que des stratégies développées lors de l'encodage du message linguistique sous forme articulatoire et acoustique. Du modèle purement géométrique de Mermelstein [2] au modèle statistique de Maeda [3], la démarche est souvent restée la même : chercher à simuler la forme du conduit vocal en agissant sur des paramètres aptes à rendre compte des effets propres à chacun des articulateurs de la parole. L'étude du processus de production de la parole s'étant étendue du domaine de compréhension des origines du signal acoustique vers des connaissances plus générales des relations entre niveau symbolique et niveau physique, les modèles articulatoires ont eu de plus en plus tendance à mimer la réalité physiologique par des approches biomécaniques. Si l'on veut, en effet, essayer de dégager des stratégies de commandes des articulateurs, et en particulier avancer sur le problème des relations entre la chaîne discrète des commandes phonémiques et les continua articulatoires et acoustiques, les modèles articulatoires doivent être aussi proches que possible du processus physiologique de production de la parole. Parce que l'être humain ne contrôle pas des paramètres mais des groupements de muscles, qui agissent en synergie, les modèles biomécaniques du conduit vocal tentent de plus en plus d'introduire une modélisation des structures musculaires.

L'axe modélisateur de l'équipe articulatoire de l'Institut de la Communication Parlée (ICP) se situe au coeur de cette politique de recherche. Un modèle biomécanique de l'ensemble mandibule + os hyoïde a ainsi été développé avec l'aide des spécialistes du contrôle moteur que sont Anatol Feldman et David Ostry [4]. Ce modèle contrôle la position de l'os hyoïde ainsi que la rotation et la translation de la mandibule grâce à un ensemble de sept muscles ou groupes de muscles. De la même façon, nous nous proposons de décrire dans ce papier, le modèle biomécanique de la langue développé à l'ICP ainsi que la façon dont nous entendons contrôler ce modèle en production de parole. Dans un premier temps, sera détaillée la façon dont le modèle a été défini. Ensuite, nous présenterons nos propositions pour un contrôle du modèle qui puisse rendre compte ① des synergies musculaires observées en parole et ② du lien entre la chaîne phonémique discrète et les représentations continues articulatoires et acoustiques. Nous essaierons ainsi de revenir sur la notion de cible et d'en proposer des corrélats au niveau des commandes motrices, espérant ainsi avancer sur les problèmes classiques d'invariance et de variabilité. Enfin, nous citerons un exemple où cette théorie du contrôle moteur est appliquée à la simulation, par le modèle, d'une transition voyelle-voyelle.

2. LE MODÈLE BIOMÉCANIQUE DE LA LANGUE

Une des propriétés fondamentales de la langue est le fait que cet articulateur est un corps déformable, pratiquement incompressible puisque majoritairement formé d'eau, dont l'évolution temporelle de la forme dépend de près d'une vingtaine de muscles internes et externes. Perkell [5] a été l'un des premiers à développer un modèle physiologique de la langue par une structure du type "matelas de ressorts", mettant ainsi en évidence le rôle de chacun des muscles sur la déformation totale de la langue. Les modèles développés, par la suite, ont intégré un outil de déformation bien mieux

adapté à la structure continue, élastique et incompressible du tissu lingual : la Méthode des Éléments Finis ([6], [7], [8]). Et ce n'est que très récemment que Wilhelms-Tricarico a "animé" un modèle 3D à base d'Éléments Finis, en fournissant aux huit muscles modélisés, des activités EMG mesurées en parole [9]. Si ces modélisations permettent de rendre compte correctement de la structure interne de la langue, elles nécessitent toutefois des temps de calcul relativement importants. Le problème devient même critique lorsqu'on cherche à intégrer la dynamique au système avec parfois plusieurs heures de calcul machine pour une seconde de mouvement. C'est une des raisons pour lesquelles nous avons adopté cette Méthode des Éléments Finis, mais en nous limitant à une description sagittale des déformations linguales. Notons que ce choix a également été motivé par le fait que de nombreuses données (cinéradiographies, enregistrements électromagnétométriques) sont disponibles dans le plan sagittal seulement.

2.1. Structure interne du modèle

Pour nous affranchir de tout problème de normalisation du locuteur dans l'évaluation acoustique de notre modélisation (cf §4 pour les simulations), nous avons choisi d'intégrer notre modèle de langue dans l'espace articulatoire d'un locuteur (PB) dont un ensemble de 1100 images cinéradiographiques ont été enregistrées à l'Institut de Phonétique de Strasbourg [10]. On a donc récupéré les contours sagittaux associés à une position de "repos" du locuteur (pas de production de parole et la langue dans sa position de repos). La radiographie correspondante nous a fourni les contours du palais, vélum, pharynx et larynx, la position et la forme de la mandibule et de l'os hyoïde, et surtout, le contour extérieur de la langue qui va servir de référence à la définition de notre modèle. Le modèle que nous proposons est décrit par une structure à base d'Éléments Finis, formée par un ensemble de 63 noeuds, délimitant 48 éléments isoparamétriques de quatre nœuds chacun. Ce choix, qui peut paraître relativement arbitraire, a tout de même été guidé par deux critères majeurs : ① la distribution interne des éléments doit refléter au mieux la structure des fibres musculaires linguales, ② sans pour autant nécessiter un nombre trop important d'éléments qui coûtent en temps de calcul. Cette structure déformable est alors "fixée" à la mandibule, ainsi qu'à l'os hyoïde, en imposant un déplacement nul des nœuds correspondants.

2.2. Modélisation des muscles linguaux

En accord avec les propositions récentes de Honda et al. [11], nous avons choisi, dans un premier temps, de réduire la modélisation des muscles linguaux aux quatre muscles qui régissent principalement la forme et la position de la langue pour la production des voyelles : le muscle Génioglosse, dont nous distinguons fonctionnellement les deux parties antérieures et postérieures, et les muscles Hyoglosse et Styloglosse. Ainsi, il nous sera possible de définir des couples agonistes/antagonistes positionnant la langue sur les deux diagonales du plan sagittal : le couple Génioglosse postérieur / Hyoglosse (GGp/HG) pour la direction frontale haute - arrière basse, et le couple Génioglosse antérieur / Styloglosse (GGa/SG) pour la direction frontale basse - arrière haute (figure 1).

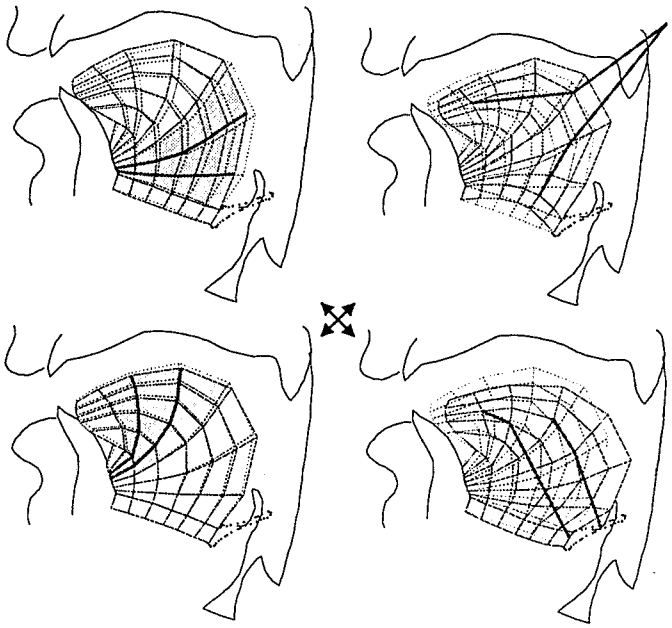


Figure 1 - Deux paires de muscles antagonistes contrôlent la position de la langue dans le conduit vocal.

3. LE CONTRÔLE DU MODÈLE

Le cadre général du modèle de langue ayant été défini, se pose alors le problème du contrôle de ce système pour la parole, et de sa relation avec le code linguistique. La recherche de régularités physiques associées à une même unité de commande linguistique s'est révélée délicate au niveau des données articulatoires et acoustiques. La rechercher au niveau des commandes musculaires individuellement semble tout autant voué à l'échec : il semble en effet difficilement concevable que nous contrôlions indépendamment chacun des vingt muscles linguaux pour produire une séquence phonologique donnée. C'est la raison pour laquelle nous optons pour un principe de contrôle qui serait capable de rendre compte des synergies musculaires gérées par le Système Nerveux Central (SNC) tout en nous permettant de poser des hypothèses sur le passage du

niveau symbolique au niveau physique. Nous nous sommes appuyés sur une théorie bien connue dans le champ des travaux sur le contrôle moteur : *l'hypothèse du point d'équilibre* proposée par Feldman, de l'Université de Montréal [12]. Feldman propose que le SNC contrôle la longueur seuil λ à partir de laquelle le recrutement des motoneurones (MNs) α commence (d'où le nom de "modèle λ" associé à cette hypothèse). L'activation musculaire dépend alors de la différence entre la longueur effective du muscle et le paramètre λ, ainsi que de la vitesse d'allongement ou de raccourcissement du muscle. Ce modèle suppose donc que les informations afférentes liées à la longueur et la vitesse du muscle viennent se superposer aux commandes efférentes des MNs α pour donner un niveau global d'activation musculaire. En suivant les propositions de Feldman, la force musculaire active est alors approximée par une fonction exponentielle estimée à partir de mesures faites sur le muscle gastrocnemius du chat [13]. Ce modèle est souvent présenté comme un *modèle de contrôle de la posture du système* : en contrôlant le seuil de recrutement λ, le SNC utilise les informations proprioceptives sur la longueur des muscles (via en particulier les fuseaux neuromusculaires). Ensuite, cet équilibre mécanique du système musculaire peut être rompu par le SNC après spécification de nouveaux seuils de recrutement des muscles impliqués dans le mouvement ; un nouvel équilibre mécanique va alors être atteint. De ce point de vue, le contrôle du mouvement revêt un caractère très particulier : le SNC spécifie les cibles (ou postures) vers lesquelles le système en mouvement évolue : la dynamique n'est alors aucunement contrôlée au niveau central mais dépend essentiellement du système périphérique. Cette hypothèse est extrêmement séduisante dans le cas de la parole puisqu'on peut y retrouver la notion de cible phonémique. Les propositions que nous formulons pour l'application à la parole de cette hypothèse du point d'équilibre sont décrites en détail dans Perrier et al. [14]. L'idée de base est que pour une chaîne phonémique donnée, une fois les paramètres prosodiques de débit et d'emphase spécifiés, le SNC associe, à chaque phonème, une cible équilibre dans l'espace λ de contrôle du système musculaire. Le passage d'une cible à l'autre, et donc la production du mouvement des articulateurs, se fait par simple transition linéaire des commandes λ. Selon les conditions prosodiques requises, le SNC va jouer sur les temps de transition d'une cible à l'autre, sur les temps de maintient d'une cible, ainsi que sur le niveau global de force mis en jeu (cocontraction). Dans des travaux récents, nous avons montré avec un modèle dynamique fonctionnel de la langue, ainsi qu'avec le modèle biomécanique de la mandibule et de l'os hyoïde de Laboissière et al. ([15], [16]), qu'un tel système de contrôle est apte à rendre compte, avec des cibles sous-jacentes invariantes, de la variabilité observée aux niveaux acoustique et articulatoire dans des conditions variables de débit et d'accentuation. Une partie de cette variabilité est planifiée au niveau central par spécification des cibles équilibres associées aux phonèmes (et fonction du contexte phonétique), tandis qu'une autre partie de cette variabilité est liée au système périphérique (et fonction des conditions prosodiques).

4. SIMULATIONS

Nous allons appliquer ici l'hypothèse du point d'équilibre à la simulation, par le modèle, d'une transition vocalique pour laquelle le mouvement de la langue est relativement simple : la transition [i-a]. Nous disposons des enregistrements cinéradiographiques du locuteur PB prononçant cette transition vocalique, et nous nous proposons de les comparer avec les signaux articulatoires et acoustiques simulés par le modèle. Nous supposons, pour cette première simulation, que les seuls muscles réellement actifs pour cette transition vocalique sont la partie postérieure du Génioglosse (surtout actif pour la voyelle [i]) et le muscle Hyoglosse (voyelle [a] surtout). D'ailleurs, comme nous le signalions plus haut, ces deux muscles agissent en synergie pour positionner la langue sur la diagonale frontale haute - arrière basse.

Selon le schéma que l'on vient de décrire, on va considérer que le mouvement résulte d'une transition linéaire des commandes λ de chacun des muscles HG et GGp entre les positions cibles $\lambda(i)$ et $\lambda(a)$ (mesurées à partir des longueurs des deux muscles sur chacune des voyelles simulées par le modèle, et après avoir retiré la composante déformation liée à la mâchoire). Le temps de transition a été adapté pour coller au mieux avec les contours de langue observés.

Le modèle de langue ayant été intégré dans un conduit vocal complet, correspondant aux contours sagittaux mesurés sur les radiographies du locuteur PB, il nous a été possible de simuler le signal acoustique par passage à la fonction d'aire. Le sonagramme obtenu, ainsi que les contours sagittaux simulés par le modèle, peuvent alors être comparés aux signaux acoustiques et articulatoires mesurés sur le locuteur PB. Cette comparaison, même si elle reste très qualitative, nous montre qu'on obtient, avec une théorie du contrôle très simplifiée, des signaux articulatoires et acoustiques tout à fait cohérents et assez proches des données originales (figure 2).

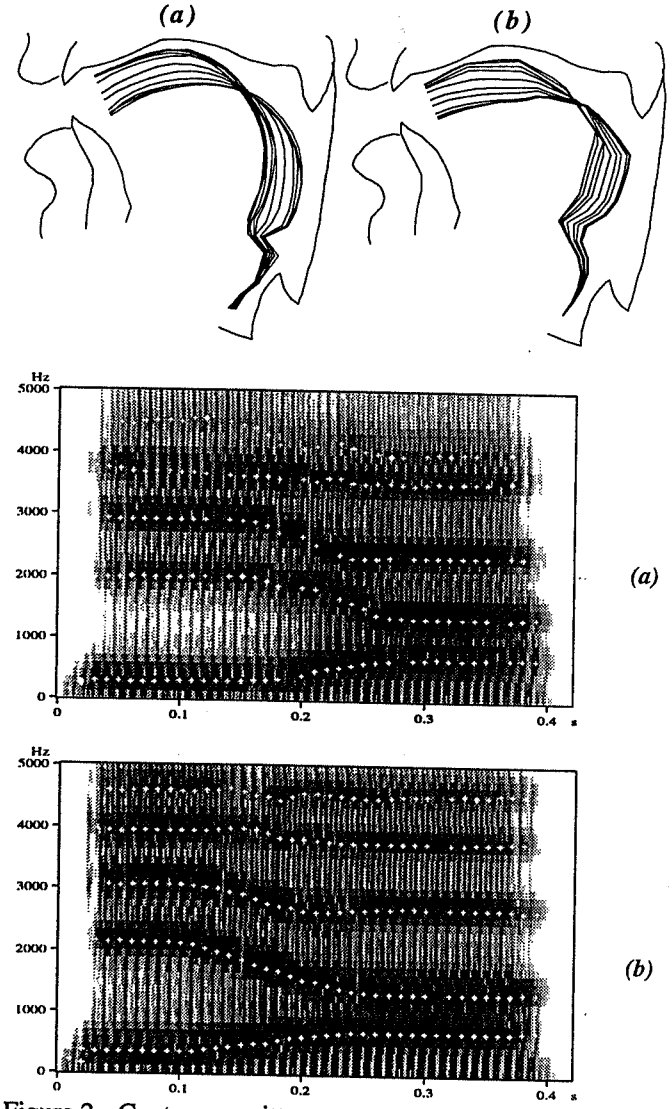


Figure 2 - Contours sagittaux et sonagramme simulés par le modèle (b) et comparés avec les données originales (a).

5. CONCLUSION

Nous avons présenté un modèle biomécanique de la langue, qui intègre les propriétés élastiques de base de cet articulateur, ainsi que les principes de génération de force d'une partie des muscles linguaux. En proposant un contrôle de ce modèle par l'hypothèse du point d'équilibre, et en revenant sur la notion de cible associée à chacun des phonèmes d'une chaîne phonémique discrète, nous avons montré que l'on pouvait générer une transition vocalique de manière cohérente dans les espaces articulatoires et acoustiques. Cette théorie du contrôle, si elle a l'avantage de poser des hypothèses simples qui alimentent le débat sur l'invariance en parole, doit bien entendu être testée sur d'autres transitions vocaliques, en impliquant d'autres synergies musculaires.

REMERCIEMENTS

Nous tenons à remercier tout particulièrement David Ostry pour toutes les discussions que nous avons pu avoir ensemble autour de ce modèle et de son contrôle. Ce travail a été supporté par l'Union Européenne (ESPRIT-BR Projet n°6975) ainsi que par la Coopération France-Québec.

RÉFERENCES

[1] **Boë L.J., Maeda S. & Perrier P. (1994).** La modélisation articulatoire : un demi siècle d'évolution entre fonctionnel, physique et biomécanique, *Actes des XXèmes Journées d'Étude sur la Parole* (p. 41-54). Trégastel, 1er au 3 Juin 1994.

[2] **Mermelstein P. (1973).** Articulatory model of speech production, *Journal of Acoust. Soc. Am.*, 53, 1070-1082.

[3] **Maeda S. (1988).** Improved articulatory model, *J. Acoust. Soc. Am.,81, S146* .

[4] **Laboissière R., Ostry D.J. & Feldman A.G. (1996)** The control of human jaw and hyoid movement, *Biological Cybernetics,* in press.

[5] **Perkell J.S. (1974).** *A physiologically-oriented model of tongue activity in speech production.* Ph.D. Thesis. Boston: Massachussetts Institute of Technology.

[6] **Kiritani S., Miyawaki K., & Fujimura O. (1976).** A Computational Model of the Tongue. *Annual Bulletin, 10* (pp. 243-252). Tokyo: Research Institute of Logopedics and Phoniatrics, University of Tokyo.

[7] **Kakita Y., Fujimura O., & Honda K. (1985).** Computation of Mapping from Muscular Contraction Patterns to Formant Patterns in Vowel Space. In V.A. Fromkin (Ed.), *Phonetic Linguistics* (pp. 133-144). Orlanda, Florida: Academic Press.

[8] **Hashimoto K. and Suga S. (1986)** Estimation of the muscular tensions of the human tongue by using a three-dimensional model of the tongue. *Journal of Acoustic Society of Japon* (E) 7,1,39-46.

[9] **Wilhelms-Tricario R. (1995)** Physiological modeling of speech production: Methods for modeling soft-tissues articulators. *J. Acoust. Soc. Am.* 97 (5), Pt.1, May 1995.

[10] **Badin P., Gabioud B., Beautemps D., Lallouache T.M., Bailly G., Maeda S., Zerling J.P. & Brock G. (1995).** Cineradiography of VCV sequences : articulatory-acoustic data for a speech production model. *15th International Congress on Acoustics,* Trondheim.

[11] **Honda K., Hirai H. & Kusakawa N. (1995).** Modeling Vocal Tract Organs Based on MRI and EMG Observations and Its Implication on Brain Function, *R.I.L.P* n°27, p 37-50.

[12] **Feldman A.G. (1966).** Functional Tuning of The Nervous System with Control of Movement or Maintenance of a Steady Posture – II Controllable Parameters of the Muscles. *Biophysics, 11,* 565-578.

[13] **Feldman A. and Orlovsky G.N. (1972).** The influence of different descending systems on the tonic reflex in the cat. *Experimental Neurology* 37:481-494.

[14] **Perrier P. Lœvenbruck H. & Payan Y. (1996).** Control of tongue movements in speech: The Equilibrium Point hypothesis perspective. *J. Phonetics,* 24, in press.

[15] **Lœvenbruck H. & Perrier P. (1993)** Vocalic reduction: prediction of acoustic and articulatory variabilities with invariant motor commands, *Proceedings of the 3rd European Conference on Speech Communication and Technology,* vol.1, (pp. 85-88), Berlin, Germany.

[16] **Lœvenbruck H., Perrier P. & Ostry D.J. (1995).** Equilibrium Point hypothesis and articulatory targets in speech: a description from simulations of empirical data using a biomechanical model of the jaw, *Proceedings of the XIIIth International Congress of Phonetic Sciences* Vol.2, pp. 462-464 Stockholm.

DETERMINATION, PAR IRM, DE L'OUVERTURE AU VELUM DES VOYELLES NASALES DU FRANCAIS

Didier DEMOLIN, Martine GEORGE, Véronique LECUIT,
Thierry METENS et Alain SOQUET

Université Libre de Bruxelles
Service de Linguistique Générale
Institut des Langues Vivantes et de Phonétique
Unité de Résonance Magnétique de l'Hopital Erasme-
Tél.: ++ 32 2 650 45 07 - Fax: ++ 32 2 650 24 50 -email: ddemoli@ulb.ac.be

ABSTRACT

MRI techniques have been used to describe velum opening for French vowels. Data were recorded with two subjects, based on 18 joined axial cuts of 4mm thickness. Differences of velum opening are calculated from areas measured in the tract determined by the velum lowering. Results show that for both subjects, the back vowel [ɔ̃] has the smallest opening while some variations are observed for the other vowels.

1 INTRODUCTION

Les techniques de résonance magnétique (IRM) ont été utilisées pour obtenir des coupes médio-sagittales du conduit vocal ainsi que pour extraire à partir de celles-ci les fonctions d'aires (Demolin et al. 1995). Les progrès récents de cette technique permettent de collecter des données avec une meilleure précision et beaucoup plus vite que précédemment. L'imagerie par résonance magnétique permet d'obtenir des coupes précises de l'ouverture du voile du palais dans différents plans, coronal, transversal ou sagittal, ceci sans aucun effet invasif pour le sujet. L'IRM vient compléter un ensemble de techniques déjà utilisées pour étudier les phénomènes de nasalisation telles que celles qui on été décrites dans Krakow et Huffman (1993).

Dans cet article, nous nous proposons de déterminer et de mesurer l'ouverture du voile du palais, pour les voyelles nasales du français. Les images ont été acquises à l'unité de résonance magnétique de l'hôpital Erasme de l'Université Libre de Bruxelles.

Les mesures ont été faites sur base de 18 coupes transversales jointes de 4 mm d'épaisseur. La matrice d'acquisition est 256*180 et le temps d'acquisition est de 15s. Le placement des coupes transversales s'est fait à partir d'une coupe médio-sagittale de la voyelle orale correspondante (cf. figures 1 et 4) prise immédiatement avant l'enregistrement des données. Le champ de vision de cette coupe est de 250mm*200mm.

L'examen du voile du palais par IRM ne pose aucun problème particulier, si ce n'est un éventuel effet de gravitation du à la position couchée du sujet pendant l'enregistrement des données. Cependant, Baer et al. (1991), Demolin et Segebarth (1992) ainsi que Demolin et al. (1995) ont montré que rien de significatif n'a pu être trouvé à ce sujet.

Les données ont été collectées avec un locuteur féminin (sujet 1) et un locuteur masculin (sujet 2) vivants à Bruxelles. La tâche des sujets était de prononcer les voyelles nasales [ɛ̃], [ɑ̃], [ɔ̃] et [œ̃] isolément. La référence était un mot qui contenait la voyelle à prononcer. Ce mot de référence était donné quelques secondes avant l'enregistrement par un des examinateurs.

2. EXAMEN DES DONNEES

Les figures 2a et 2b donnent les coupes les plus significatives obtenues pour les voyelles [ɛ̃] et [ɑ̃] avec le sujet 1. Les coupes 1, 2, 3, montrent la séparation du conduit oral et de l'ouverture nasale, les tissus entre les deux zones sombres (l'ouverture nasale vers le bas et le conduit oral vers le haut) étant ceux du voile du palais. Les coupes situées en deçà donnent les aires au niveau du pharynx, celles qui sont au delà se situent dans les fosses nasales ou dans une région non-significative pour notre étude. La coupe médio-sagittale de la figure 1 ([ɑ̃]) montrent assez clairement que l'extrémité de la luette semble toucher le dos de la langue, ce fait est plus marqué encore avec la voyelle postérieure mi-ouverte [ɔ̃].

Les figures 4a et 4b donnent les coupes les plus significatives obtenues pour les voyelles [ɛ̃] et [ɑ̃] avec le sujet 2. Les coupes 1, 2, 3, montrent la séparation du conduit oral et de l'ouverture nasale, les tissus entre les deux zones sombres (l'ouverture nasale vers le bas et le conduit oral vers le haut) étant ceux du

voile du palais. Les coupes situées en deçà donnent les aires au niveau du pharynx, celles qui sont au delà se situent dans les fosses nasales ou dans une région non-significative pour notre étude.

L'examen des images qui sont présentées aux figures 2a et 2b, permet de constater les différences intéressantes qui existent entre la voyelle antérieure [ɛ̃] et la voyelle postérieure [ɑ̃] chez le sujet 1. A la figure 2a ([ɛ̃]), on peut voir que ni le voile, ni la luette ne sont en contact avec la langue. A la figure 2b, il apparaît clairement que la luette touche le creux sagittal de la langue à la coupe 1, en laissant un passage libre de part et d'autre du contact qui apparaît presque complet à la coupe 2. Cette différence dans la position de la luette entre les voyelles postérieures et antérieures se retrouve avec les autres voyelles que nous avons examinées.

L'examen des images qui sont présentées aux figures 4a et 4b permet de constater les différences intéressantes qui existent entre la voyelle antérieure [ɛ̃] et la voyelle postérieure [ɑ̃] chez le sujet 2. A la figure 4a ([ɛ̃]), on peut voir que ni le voile, ni la luette ne sont en contact avec la langue. A la figure 4b, il apparaît clairement que la luette touche le creux sagittal de la langue à la coupe 1, en laissant un passage libre de part et d'autre du contact qui apparaît presque complet à la coupe 2. Cette différence dans la position de la luette entre les voyelles postérieures et antérieures se retrouve avec les autres voyelles que nous avons examinées.

3. DISCUSSION

A partir des images obtenues, et pour chacune des voyelles nasales, les aires correspondant aux coupes 1 à 10 (sujet 1) et 6 à 15 ont été dessinées (sujet 2) (figure 7 et 8). Les aires des coupes 5, 6, 7, 8, 9 ont ensuite été mesurées à l'aide d'une méthode décrite dans Lecuit (1995) pour le sujet 1. Chaque aire a été mesurée 3 fois. La moyenne de ces mesures est donnée en cm^2, à la table 1. Pour le sujet 2, les aires mesurées et présentées à la table 2 sont les aires 11 à 14.

Pour le sujet 1, les résultats montrent que pour chaque voyelle, l'aire 6 est la plus petite. L'aire 5 qui est en dessous de l'aire 6, est plus grande dans tous les cas. Au dessus de l'aire 6, on constate un accroissement progressif des aires examinées, dont l'ordre de grandeur varie différement pour chacune des voyelles. On peut donc considérer, sur base de ces documents et pour ce sujet, que c'est l'aire 6 qui donne la mesure de l'ouverture du voile. L'ordre de grandeur entre les voyelles examinées à l'aire 6 est le suivant: ɔ̃>ɑ̃>ɛ̃>. Les deux voyelles postérieures [ɔ̃] et [ɑ̃] ont l'ouverture la plus petite et les deux voyelles antérieures [ɛ̃] et [œ̃] ont l'ouverture la plus grande. La plus petite ouverture de la voyelle ɔ̃ est une conséquence de la position la plus haute et la plus postérieure de la masse de la langue, dans les voyelles que nous avons examinées. L'ordre de grandeur constaté reflète bien les différences articulatoires entre les voyelles. L'aire 6 de [ɑ̃] qui est plus grande que celle de [ɔ̃] est également celle d'une voyelle postérieure, mais dont le degré d'aperture est plus grand. On constate moins de différences entre les deux voyelles antérieures [ɛ̃] et [œ̃] dont la différence articulatoire réside dans une centralisation plus grande et dans l'arrondissement de [œ̃]. L'aire 6 de [ɛ̃] est légèrement plus petite que celle de [œ̃]. Il ne semble toutefois pas possible de montrer, sur la base de nos données, que ce soit lié à la différence d'articulation entre les deux voyelles.

Pour le sujet 2, les résultats montrent que pour chaque voyelle, l'aire 12 est la plus petite. L'aire 11 qui est en dessous de l'aire 12, est plus grande dans tous les cas. Au dessus de l'aire 12, on constate un accroissement progressif des aires examinées, dont l'ordre de grandeur varie différemment pour chacune des voyelles. On peut donc considérer, sur base de ces documents et pour ce sujet, que c'est l'aire 12 qui donne la mesure de l'ouverture du voile. L'ordre de grandeur entre les voyelles examinées à l'aire 12 est le suivant: ɔ̃>ɑ̃>œ̃>ɛ̃. Les deux voyelles [ɔ̃] et [œ̃] ont l'ouverture la plus petite et les deux voyelles [ɑ̃] et [ɛ̃] ont l'ouverture la plus grande. La plus petite ouverture de la voyelle [ɔ̃] est une conséquence de la position la plus haute et la plus postérieure de la masse de la langue, dans les voyelles que nous avons examinées. L'ordre de grandeur constaté reflète bien les différences articulatoires entre les voyelles. L'aire 12 de [ɑ̃] qui est plus grande que celle de [ɔ̃] est également celle d'une voyelle postérieure, mais dont le degré d'aperture est plus grand. On constate moins de différences entre les deux voyelles antérieures [ɛ̃] et [œ̃] dont la différence articulatoire réside dans une centralisation plus grande et dans l'arrondissement de [œ̃].

4. CONCLUSION

Ces résultats, qui doivent être confirmés par des études sur d'autres sujets, permettent de penser qu'il existe des différences plus subtiles entre les voyelles nasales que ce qui est généralement décrit. Vaissière (1995), dans une discussion synthétisant les données sur les

différences intrinsèques de hauteur du voile du palais entre les sons, affirme que l'activité du levator palatini est plus importante (et le voile plus élevé) pour les voyelles fermées que pour les voyelles ouvertes. Les voyelles nasales du français étant toutes plutôt ouvertes (entre ouvert et mi-ouvert), les résultats obtenus dans notre étude permettent de penser qu'il y aurait une différence pertinente d'activité du levator palatini entre le niveau ouvert [ɔ̃], [ɑ̃] et le niveau mi-ouvert [ɛ̃], [œ̃]. L'étude sur des langues qui ont des voyelles nasales à différents niveaux d'aperture (de fermé à ouvert) permettra peut-être de montrer si cette hypothèse est pertinente et si la position postérieure de la masse de la langue à une influence sur le degré d'ouvrture du voile du palais.

BIBLIOGRAPHIE

Baer, T., J.C. Goore, L. C. Gracco and P. W. Nye. 1991. Analysis of vocal tract shape and dimensions using magnetic resonance imaging: vowels. *J.A.S.A.* 90. 799-828.

Demolin, D., J-M Hombert, V. Lecuit, A. Soquet et C. Segebarth. 1995. An MRI study of French vowels. *Eurospeech*, Madrid, pp. 2235-2238.

Demolin, D. & Segebarth, C. 1992. Analyse de production de voyelles de quelques langues du Soudan central par IRM." *Actes des XIXèmes Journées d'Etude de la Parole*, Université Libre de Bruxelles.

Krakow M. K., et R. A. Huffman. 1993. *Nasals, Nazalization and the Velum.* Academic Press. *New-York.*

Lecuit, V. 1995. *Sagittal Cut to Area Function Transformations: A Comparative Study.* Mémoire de Licence. Université Libre de Bruxelles.

Vaissière, J. 1995. Nasalité et phonétique. *Voile pathologique*. Société Française de Phoniatrie et Société Française d'Acoustique. 81-92.

Table 1: Moyenne des sections des coupes 5 à 9 pour les quatre voyelles nasales (S1)

Coupe	[]	[]	[]	[]
5	1.09	1.05	0.71	0.58
6	0.90	0.84	0.66	0.56
7	1.33	1.05	1.10	0.98
8	1.62	1.51	1.31	1.10

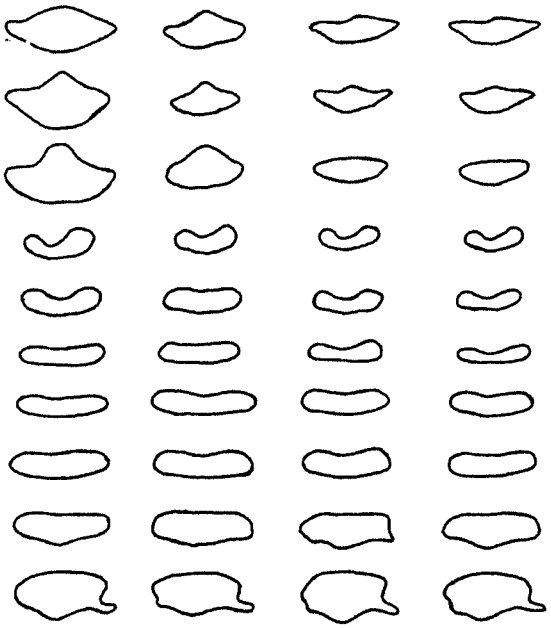


Figure 5: Sections des 4 voyelles nasales S1 (coupes 1 à 10)

Table 2: Moyenne des sections des coupes 11 à 14 pour les quatre voyelles nasales (S2)

Coupe	[œ̃]	[ɛ̃]	[ɑ̃]	[ɔ̃]
11	3.74	2.04	1.66	1.17
12	1.44	1.73	1.70	1.10
13	1.93	2.23	1.89	1.08
14	1.70	2.12	2.05	1.32

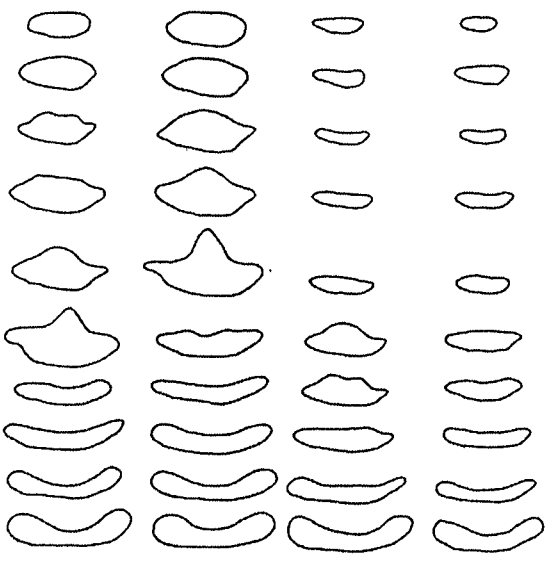


Figure 6: Sections des 4 voyelles nasales S2 (coupes 6 à 15)

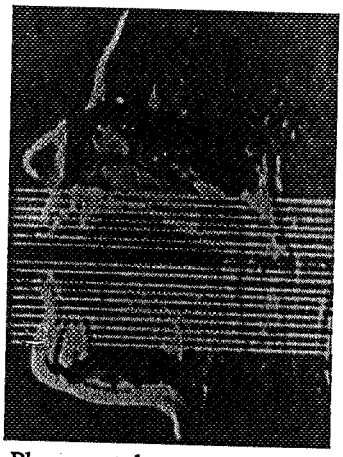


Figure 1: Placement des coupes transversales [ɑ̃]

Figure 3: Placement des coupes transversales [ɑ̃]

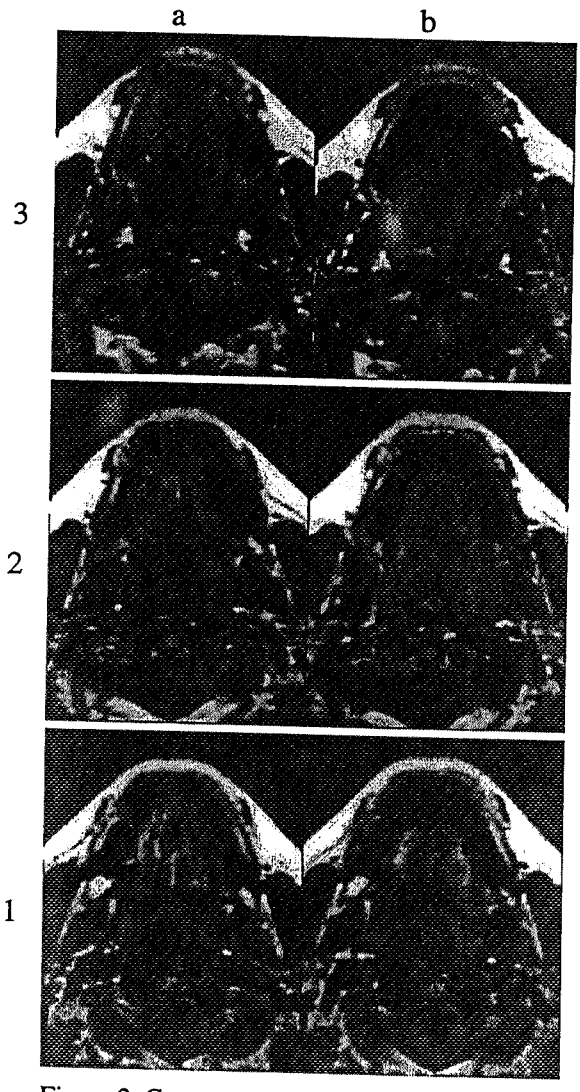


Figure 2: Coupes transversales [ɛ̃] et [ɑ̃] S1

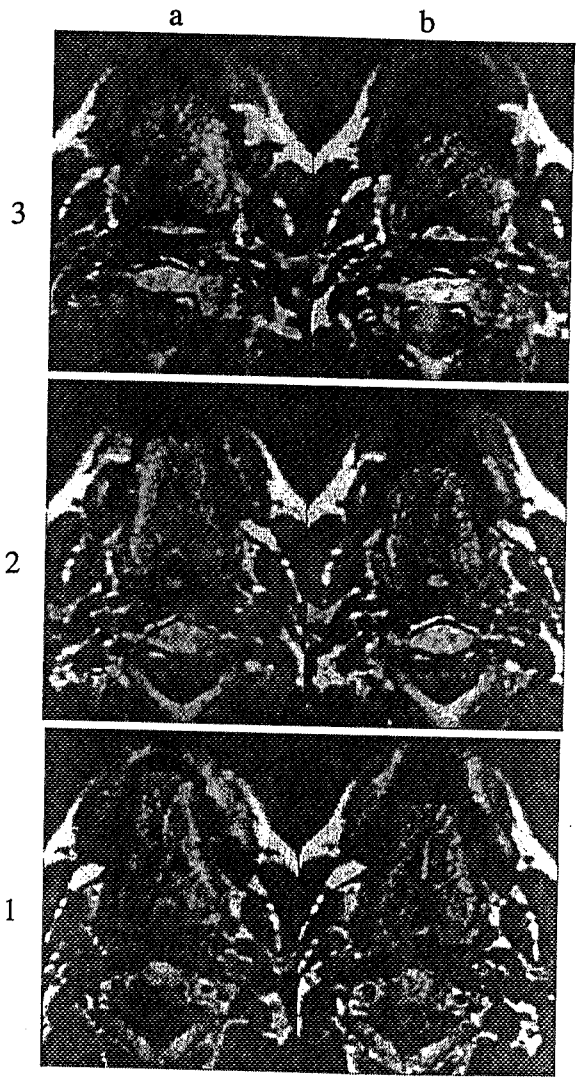


Figure 4: Coupes transversales [ɛ̃] et [ã] S2

ÉMERGENCE DE PROTOTYPES SENSORI-MOTEURS À PARTIR D'EXEMPLAIRES AUDIO-VISUELS

Gérard BAILLY
Institut de la Communication Parlée - URA CNRS 368 - INPG & Université Stendhal
46 avenue Félix Viallet, 38031 Grenoble Cedex 1, France
E-Mail : bailly@icp.grenet.fr

ABSTRACT

This paper deals with automatic learning of the control parameters for an articulatory speech synthesizer with 10 degrees of freedom. The control model is quickly presented. We demonstrate here how the optimal control space for each class of sounds can emerge from a massive inversion of audio-visual stimuli. The set of proprioceptive and exteroceptive inversion results have been compared using a Canonical Discriminant Analysis. The close observation of the topology and discrimination performance of the sensori-motor "maps" produced for each class of sounds demonstrates that the most efficient control space for vowels is the acoustic space and the geometric space for occlusives. The choice for vowels is corroborated by the simulated kinematics of some vowel-vowel transitions.

1. INTRODUCTION

Divers modèles de production de parole ont été proposés dans la littérature allant du simple modèle source-conduit à une modélisation de plus en plus proche de la réalité physique. Le degré de finesse de description de la géométrie s'accompagne d'un jeu de plus en plus grand de paramètres de contrôle: le modèle de Fant piloté par 3 paramètres (aire aux lèvres Al, position Xc et aire Ac de la constriction linguale) décrit les paramètres géométriques essentiels pour le contrôle acoustique des voyelles; les modèles articulatoires proposés par Mermelstein et Maeda [9] possèdent une dizaine de paramètres de contrôle alors que les modèles de contrôle musculaires d'articulateurs rigides et d'hydrostats [13] comportent typiquement plusieurs dizaines de variables à contrôler.

Le contrôle de tels systèmes à forts degrés de liberté en excès ne peut dès lors être envisagé par des techniques classiques de synthèse par règles. On doit donc dissocier l'espace d'encodage des tâches phonologiques - paramétrisation distale ou encodage des buts - de l'espace - proximal - de contrôle du modèle de production (cf. Fig.1).

Parmi les modèles de contrôle, on peut citer le modèle de dynamique gestuelle [14] qui suppose un encodage géométrique des tâches. Ce modèle suppose que la coarticulation est gérée au niveau distal par un modèle d'anticipation et de superposition. Une version allégée de ce modèle proposée par Honda & Kaburagi [6] suppose que la partition n'impose que des points de rendez-vous. Dans ce type de système, l'espace distal est le même quelque soit la tâche alors que Stevens & Bickley [16] propose un espace composite où les paramètres de contrôle de haut niveau comportent aussi bien des paramètres acoustiques (formants) que des variables géométriques et aérodynamiques.

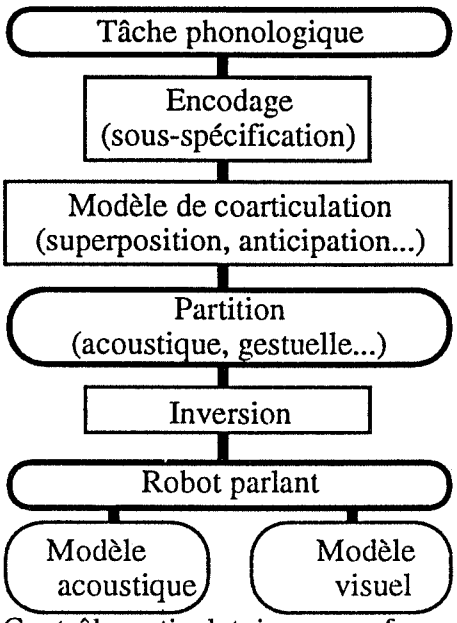


Fig.1: Contrôle articulatoire par formation de trajectoires. Le module d'inversion sélectionne à chaque instant le geste le plus approprié et le moins coûteux.

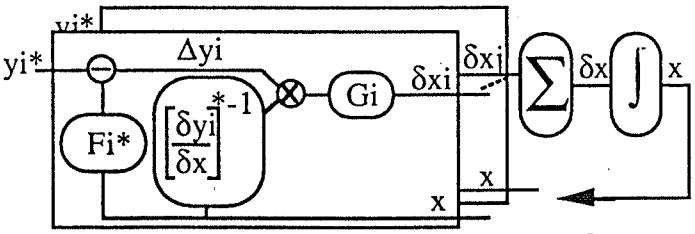


Fig.2: Contrôle composite par rétroaction. Chaque cible yn* est un point d'équilibre sur une carte spécifique (acoustique ou autre). x est le vecteur d'articulateurs courants résultants de l'intégration temporelle de la somme des déplacements élémentaires δxn. Fn* est la fonction réalisant la transformation directe. Le contrôle en boucle sur un espace est obtenu par le jacobien inverse.

2. CONTRÔLE PAR CHAMPS DE FORCE COMPOSITES SUR CARTES SENSORI-MOTRICES

Le modèle de contrôle que nous allons paramétrer est issu du projet esprit SpeechMaps. L'articulotron est bâti sur trois principes (Fig.2):

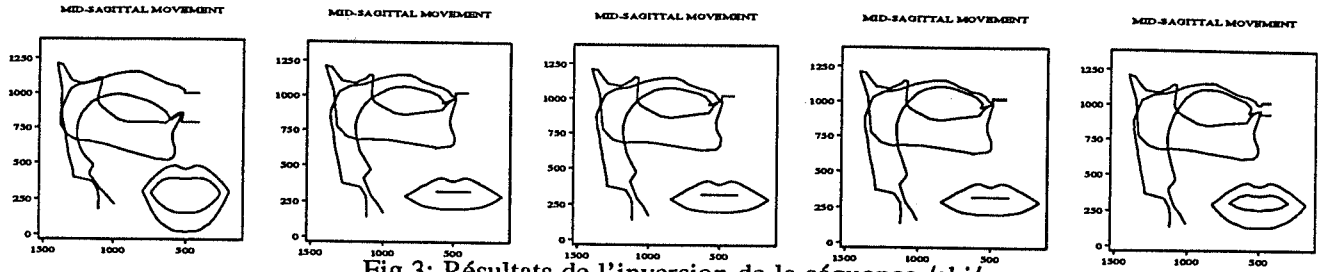


Fig.3: Résultats de l'inversion de la séquence /abi/.

• un codage positionnel des cibles; chaque cible - typiquement une ou plusieurs gaussiennes [10] - positionnée dans un espace de représentation quelconque de la parole génère un champ de force qui attire le point représentant la trame courante vers cette cible.

• un rétro-propagation au niveau proximal; la force d'attraction ainsi créée est traduite en une force articulatoire par inversion du jacobien de la transformation proximale-distale.

• un codage composite et compositionnel; chaque cible possède une fonction d'émergence qui peut chevaucher d'autres fonctions. Les champs de force ainsi créé s'additionnent dans chaque niveau de représentation et le mouvement ainsi créé génère une trajectoire cohérente dans tous les espaces de représentation.

Nous proposons ici un modèle d'apprentissage automatique du codage positionnel le plus efficace. Ce modèle est testé sur un ensemble de réalisations naturelles.

3. CARACTÉRISATION ET INVERSION D'EXEMPLAIRES AUDIOVISUELS

3.1. Le corpus

Il comprend 78 stimuli de type V1CV2 où V est une des 10 voyelles du français et C une occlusive voisée. Ces stimuli ont été prononcés par le locuteur pour lequel un modèle articulatoire est disponible et un modèle acoustique a été adapté [1].

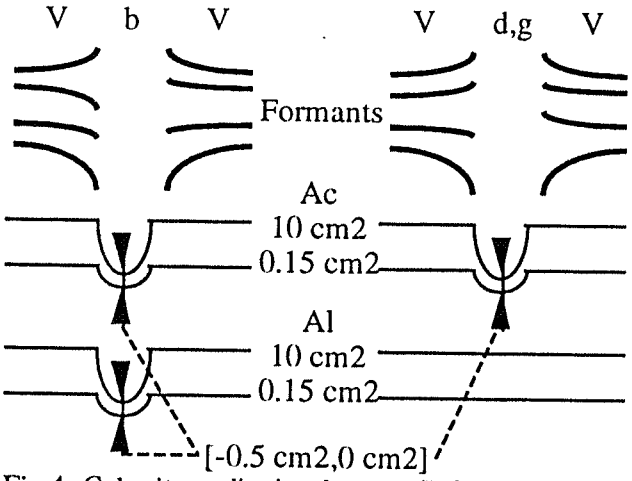


Fig.4: Gabarits audiovisuels pour /VCV/

3.2. Caractérisation audiovisuelle

Les stimuli sont caractérisés par l'évolution des quatre premiers formants (ces formants ont été déterminés semi-automatiquement) et d'estimés grossiers de l'aire aux lèvres. Ces valeurs sont définies sur des régions délimitées par les frontières vocaliques (cf. Fig.4) :

• les formants peuvent avoir une valeur quelconque dans l'occlusion

• Ac doit être négative ou nulle au centre de l'occlusion et inférieure à 0.1 cm2 aux frontières vocaliques.

• le gabarit de Al est supposé évoluer comme Ac pour la bilabiale alors que Al doit être supérieure à 0.1 cm2 pour /d,g/.

3.3. Modèle d'inversion

Les transformations articulatoire-géométrique et articulatoire-acoustique ont été estimées par une régression multinomiale d'ordre 4 sur un dictionnaire de 16000 formes. Ce dictionnaire comporte les trames de rayons-X utilisées pour construire le modèle articulatoire et une inspection aléatoire de l'espace maximal articulatoire en éliminant les doubles constrictions et les aires de constriction estimées inférieures à -0.1 cm2[1].

3.4. Inversion

L'inversion est itérative et procède d'une descente de gradient composite (addition des inverses des dérivées partielles) sous contrainte de minimisation d'accélération[2]. L'exemple d'inversion montré Fig.3 pour la séquence /abi/ montre comment le système d'inversion est conduit à anticiper l'avancée-montée de la langue du /i/ dans la réalisation de l'occlusive bilabiale.

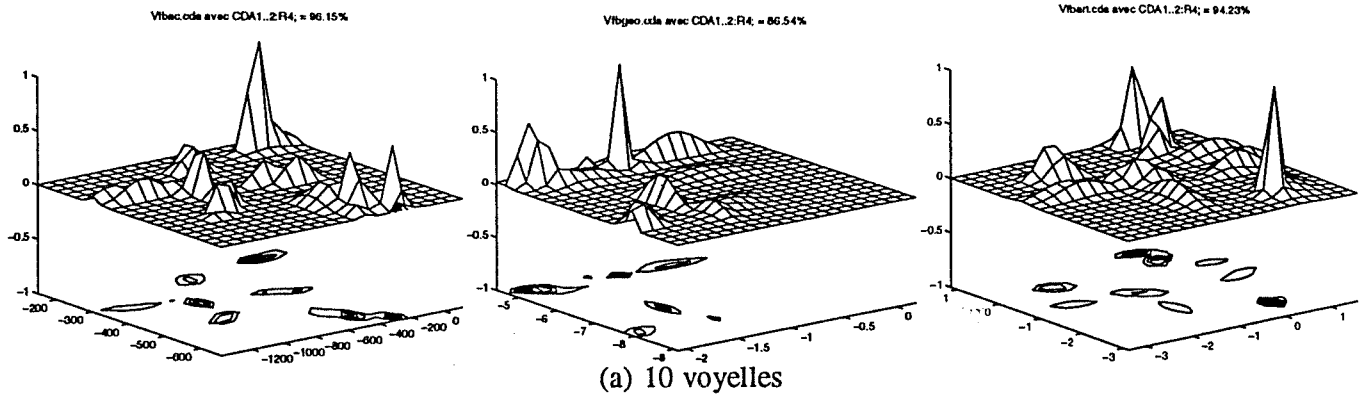
Chaque stimuli articulatoire a été synthétisé par modélisation dynamique temporelle du conduit vocal en utilisant un modèle élaboré de la source vocale [12]. Tous les stimuli identifiés par deux experts ont été considérés comme corrects. Seuls 5 stimuli ont donc été écartés dont 3 étaient prononcés en contexte fermé (V1,V2 dans /u,o/).

4. CARACTÉRISATION SENSORI-MOTRICE

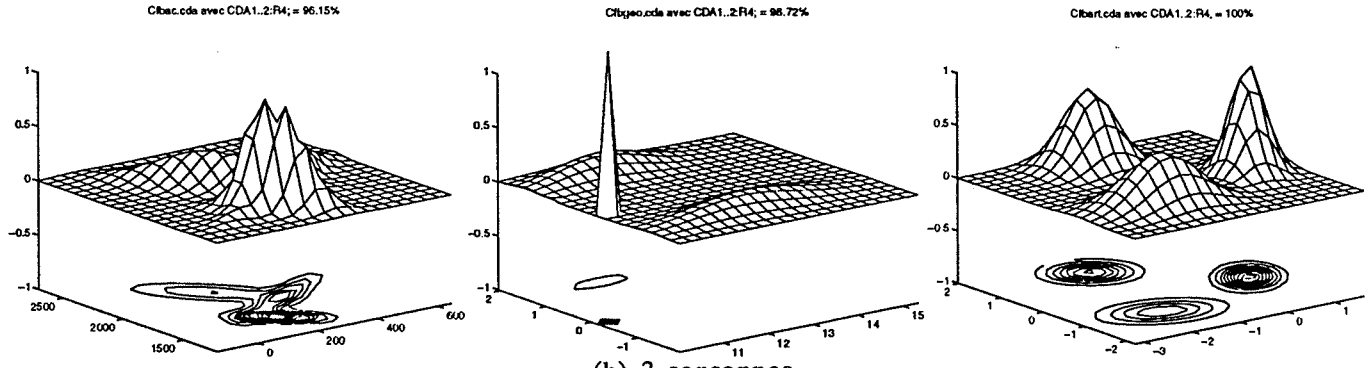
Lorsque chaque stimuli externe a bien été imité par le robot, il est caractérisé par un ensemble

[1]Le modèle peut générer des aires de constriction négatives afin de réaliser des constrictions plus ou moins marquées. Ceci revient à placer la cible géométrique au delà du palais, du pharynx ou de la lèvre supérieure.

[2]Cette contrainte s'efface au cours de la descente de manière similaire à Jordan [7].



(a) 10 voyelles



(b) 3 consonnes

Fig.5: Premiers plans discriminants. De gauche à droite: acoustique, géométrique et articulatoire.

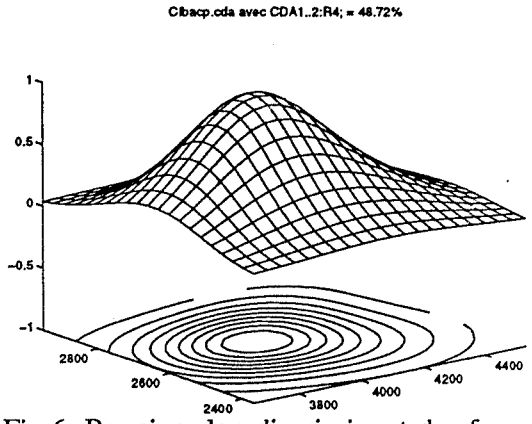


Fig.6: Premier plan discriminant des formants mesurés à l'onset vocalique.

d'informations sensori-motrices prélevées à des instants-clés. Les instants clés sont :

- Cibles vocaliques (points de stabilité spectrale maximale)
- Cibles consonantiques (centre d'occlusion)
- Débuts et fins vocaliques

Les informations sensori-motrices sont:

- Acoustique: formants estimés[3]
- Géométrique: Xc,Ac,Al
- Articulatoire: paramètres estimés

A cet ensemble de résultats d'inversion, sont ajoutés les caractéristiques mesurées des seuls 12 VCVs issus de la bases données rayons-X originale. Nous avons donc les caractéristiques sensori-motrices 72-5+12 = 79 consonnes et 158 voyelles.

Les cibles sont définies comme des régions compactes de l'espace sensori-moteur. De plus, nous construisons des topologies séparées pour consonnes et voyelles [11].

Ainsi que mentionné section 2, nous allons limiter les cibles à des attracteurs ponctuels. Bien que les mixtures de gaussiennes peuvent constituer des régions complexes connexes ou non, le champ de force généré peut présenter trop de points singuliers.

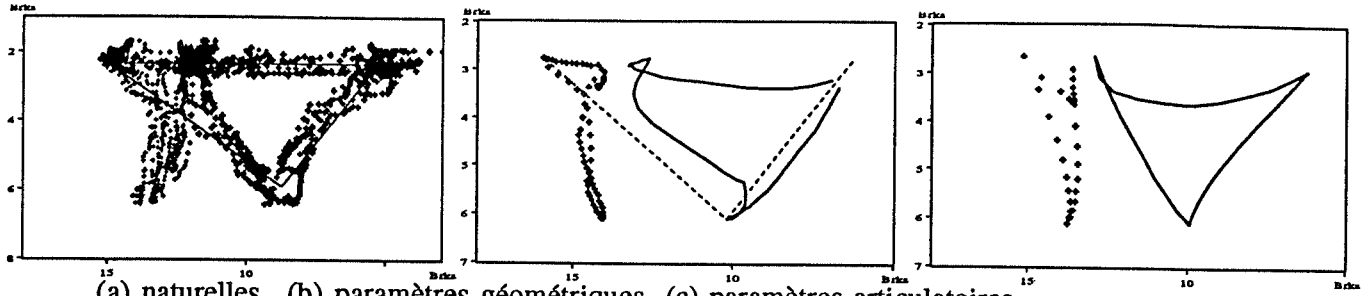
Une transformation linéaire des trois sous-espaces acoustique, géométrique et articulatoire a été effectuée de manière à produire des cibles compactes (distances intra-classe minimales) avec un contraste maximal (distances inter-classe maximales). Une Analyse Discriminante Canonique des cibles a été effectuée comme suggéré dans [15]. Les échantillons vocaliques et consonantiques ont alors été projetés dans les premiers plans discriminants de leurs trois sous-espaces respectifs (voir Fig.5).

L'examen des structures des projection et des scores d'identification démontre que les voyelles sont mieux définies dans l'espace acoustique et que les occlusives le sont dans l'espace géométrique. La faible rentabilité de l'espace acoustique pour les occlusives est démontrée Fig.6: lorsque l'information acoustique n'est pas prélevée à la cible acoustique estimée mais à l'onset vocalique, la discrimination est juste au-dessus de la chance.

5. Contrôle des voyelles

Le modèle de prédiction des systèmes

[3]Les formants des occlusives sont estimés en ouvrant légèrement (aire minimale dans le conduit 0.0001 cm2)



(a) naturelles (b) paramètres géométriques (c) paramètres articulatoires

Fig.7: Trajectoires acoustiques maximales (a) naturelles; (b) et (c) produites par interpolation linéaire dans le plan F1-F2 (traits pleins) superposées à celles dans le plan F1-F3 (croix) .

vocaliques le plus efficace [8,4] exploite un critère de dispersion maximale sur l'espace acoustique. Cet espace se révèle être aussi le plus discriminant. Mais est-il le plus apte au contrôle de la cinématique des trajectoires.

Notre modèle de contrôle est basé sur une modulation de champs gaussiens sur des cartes sensori-motrices. Si ces cartes sont connexes[4], la trajectoire produite par une modulation barycentrique de deux attracteurs ponctuels sera rectiligne. Examinons le cas de transitions voyelle-voyelle pour deux types de robot: (a) un modèle de constrictions [5]; (b) le modèle articulatoire. Les trajectoires maximales (/ai/, /iu/ et /ua/) sont réalisées en interpolant linéairement les paramètres pour les voyelles cardinales proposées respectivement dans [5] et [2]. Les trajectoires formantiques résultantes sont présentées Fig.7 et peuvent être comparées avec les trajectoires acoustiques réelles. On voit que ces dernières sont pratiquement rectilignes dans l'espace R1-R2 [3] alors que les trajectoires simulées présentent des non-linéarités graves : voir notamment la trajectoire /ai/ pour Fant et la centralisation de la trajectoire /ui/ pour le modèle articulatoire.

CONCLUSIONS

Nous avons montré que des représentations sensori-motrices adaptées peuvent émerger par accumulation d'inversions. La partition gestuelle peut donc s'enrichir par expérience d'un encodage des tâches de plus en plus adapté et efficace.

Il reste à apprendre d'autres sons et à chercher à apprendre automatiquement la synchronisation des modulations des champs de force dans les divers sous-espaces pour assurer une première version d'un système de contrôle articulatoire entièrement basé sur l'expérience.

BIBLIOGRAPHIE

[1] Badin, P., Gabioud, B., Beautemps, D., Lallouache, T., Bailly, G., Maeda, S., Zerling, J.P., and Brock, G. Cineradiography of vcv sequences: articulatory-acoustic data for a speech production model, *ICA*, 349-352, 1995.

[2] Bailly, G., Boë, L.J., Vallée, N., and Badin, P. Articulatori-acoustic prototypes for speech production, *EUROSPEECH*, 2:1913-1916, 1995.

[3] Bailly, G. Caracterisation of formant trajectories by tracking vocal tract resonances, In Beekmans, R., Jospa, P., Schoentgen, J., and Serniclaes, W., editors, *Levels in speech communication : relations and interactions*, 91-102. Elsevier, Amsterdam, 1995.

[4] Boë, L.J., Schwartz, J.L., and Vallée, N. The prediction of vowel systems: perceptual contrast and stability, In Keller, E., editor, *Fundamentals of speech synthesis and speech recognition*, 185-214. John Wiley and Sons, Chichester, 1994.

[5] Fant, G. Vocal tract area functions of swedish vowels and a new three-parameter model, *ICSLP*, 1:807-810, 1992.

[6] Honda, M. and Kaburagi, T. A dynamical articulatory model using potential task representation, *ICSLP*, 1:179-184, 1994.

[7] Jordan, M.I. Motor learning and the degrees of freedom problem, In Jeannerod, M., editor, *Attention and Performance XIII*, Lawrence Erlbaum, Hillsdale, NJ, 1990.

[8] Liljencrants, J. and Lindblom, B. Numerical simulation of vowel quality systems: The role of perceptual contrasts, *Language*, 48:839-861, 1972.

[9] Maeda, S. Improved articulatory model. *JASA*, 81(S1):S146, 1988.

[10] Morasso, P. and Sanguinetti, V. Self-organizing topographic maps and motor planning. In Cliff, D., Husbands, P., Meyer, J., and Wilson, S., editors, *From animals to animats* 3, 214-220. MIT Press, Cambridge, MA, 1994.

[11] Öhman, S.E.G. Numerical model of coarticulation, *JASA*, 41:310-320, 1967.

[12] Pelorson, X., Hirschberg, A., Wijnands, A., Bailliet, H., Vescovi, C., and Castelli, E. Description of the flow through the vocal cords during phonation. application to voiced sounds synthesis, *Acta Acustica*, (to appear).

[13] Perkell, J. *Physiology of Speech Production*, MIT Press, Cambridge, MA, 1969.

[14] Saltzman, E.L. and Munhall, K.G. A dynamical approach to gestural patterning in speech production *Ecological Psychology*, 1(4):1615-1623, 1989.

[15] Soquet, A. and Saerens, M. A comparison of different acoustic and articulatory representations for the determination of place of articulation of plosives, *ICSLP*, 2:1643-1646, 1994.

[16] Stevens, K.N. and Bickley, C.A. Constraints among parameters simplify control of klatt formant synthesizer. *J. of Phonetics*, 19:161-174, 1991.

[4]On peut aller d'un point à un autre sans sortir de la carte et les autres cartes ne limitent pas physiquement la trajectoire : les articulateurs peuvent "suivre".

ESTIMATION DE TRAJECTOIRES ARTICULATOIRES À PARTIR DE TRANSITIONS FORMANTIQUES : APPLICATION À L'ANALYSE DE SÉQUENCES V_1V_2 ET V_1CV_2

Alain Soquet[1] et Martine George[2]

Laboratoire de Phonétique Expérimentale – Université Libre de Bruxelles
50 av. F.-D. Roosevelt, CP110, B-1050 Bruxelles– Belgique
Tél.: (32 2) 650 20 18 – Fax: (32 2) 650 20 07 – e-mail asoquet@ulb.ac.be
[1]Projet de la Communauté Française de Belgique, ARC 93/98-168 – [2]Bourse U. L. B.

ABSTRACT

In previous works, we have shown that connectionnist network properly trained are able to determine articulatory parameters from the first three formant frequencies. This method does not take into account the contextual acoustic and articulatory information. In order to overcome this limitation, we propose to model the speech dynamic of each articulator with a set of sigmoïds. The analysis consists in determining the sigmoïd parameters that satisfy an acoustic criterion. This method allows to analyse V_1V_2 and V_1CV_2 logatomes (where C is a plosive and V_1 and V_2 are French oral vowels). The results show a good adequacy between measured formants and those obtained with the articulatoy modeling. At the articulatory level, both V_1V_2 and V_1CV_2 transitions are realistic and interpretable.

1. INTRODUCTION

Dans des travaux précédents (Jospa et al. [2], George et al. [1], Soquet [6]), nous avons utilisé un réseau connexionniste pour déterminer les paramètres de commande de modèles du conduit vocal à partir des valeurs instantanées des trois premiers formants. Cette méthode s'est montrée capable de fournir des configurations articulatoires réalistes pour les sons vocaliques. Cependant, elle ne tient compte d'aucune information contextuelle relative aux configurations acoustiques et/ou articulatoires passées ou futures. Pour remédier à cette limitation, nous proposons de prendre en compte le caractère dynamique de la parole en modélisant le mouvement de chacun des articulateurs par un ensemble de sigmoïdes.

2. MODÉLISATION

De manière générale, le système de production permet, à partir d'un ensemble de commandes articulatoires, de produire un signal acoustique.

Dans le cadre de cette étude, nous considérerons que le système de production peut être décomposé en deux parties : **(i)** un modèle articulatoire qui, à partir des commandes articulatoires, décrit la géométrie de la cavité orale par la fonction d'aire et **(ii)** un modèle acoustique qui simule la physique de la propagation du son dans le conduit vocal et permet le calcul d'indices acoustiques à partir de la fonction d'aire.

Des études de trajectoires articulatoires V_1V_2 indiquent que ces trajectoires sont simples mais non-linéaires (George et al. [1], Kaburagi et Honda [3]). Nous proposons de caractériser le mouvement de chacun des paramètres de commande articulatoire par un ensemble de sigmoïdes. En effet, celles-ci permettent de décrire simplement les trajectoires observées.

Le modèle sigmoïdal utilisé pour l'étude des séquences V_1V_2 (où V_1 et V_2 sont des voyelles orales du Français) est décrit par l'équation (1) :

$$u_i(t) = S_{i1} + \frac{(S_{i2} - S_{i1})}{1 + e^{-p_{i0}(t - t_{i0})}} \qquad (1)$$

Ce modèle permet d'exprimer l'évolution dans le temps du $i^{ème}$ paramètre de commande articulatoire $u_i(t)$ en termes de deux parties stables et d'une transition d'une partie stable à l'autre. Les parties stables sont respectivement caractérisées par les deux valeurs asymptotiques de la sigmoïde S_{i1} et S_{i2}, et la transition par la pente p_{i0} et le point d'inflexion t_{i0}. La figure 1 illustre ce modèle.

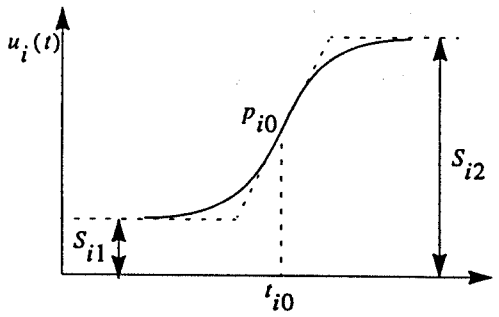


Figure 1 : Modélisation de séquences V_1V_2 .

Il serait intéressant d'étendre ce modèle d'analyse à des séquences plus complexes pour lesquelles le système ne dispose pas à tout instant d'indices acoustiques : par exemple, dans le cas des occlusives, les informations relatives au lieu d'articulation sont concentrées dans les transitions vocaliques entourant l'occlusion et le bruit d'explosion (Serniclaes [5]). Il apparaît donc clairement que durant une partie importante de l'occlusive, les indices acoustiques seuls ne permettent pas à une méthode d'analyse travaillant de manière statique de récupérer les positions des articulateurs. Nous proposons d'étendre le modèle d'analyse de V_1V_2, décrit ci-dessus, au cas de l'analyse de segments V_1CV_2 (où C est une occlusive). Dans ce cadre, nous pourrons tester l'hypothèse selon laquelle les trajectoires des articulateurs peuvent être extrapolées à partir des seules informations acoustiques présentes dans les transitions vocaliques.

Pour ce faire, nous proposons de modéliser l'évolution dans le temps des commandes articulatoires à l'aide d'une somme de deux sigmoïdes. La figure 2 illustre les paramètres de ce modèle. Les parties stables sont caractérisées par S_{i1} et S_{i2} et la transition par les deux pentes p_{i1} et p_{i2} et les deux points d'inflexion t_{i1} et t_{i2} (voir équation (2)).

$$u_i(t) = \frac{S_{i1}}{1+e^{p_{i1}(t-t_{i1})}} + \frac{S_{i2}}{1+e^{-p_{i2}(t-t_{i2})}} \quad (2)$$

Ce modèle permet de tirer directement parti des informations acoustiques disponibles à savoir les parties stables des voyelles (V_1 et V_2) adjacentes à la consonne.

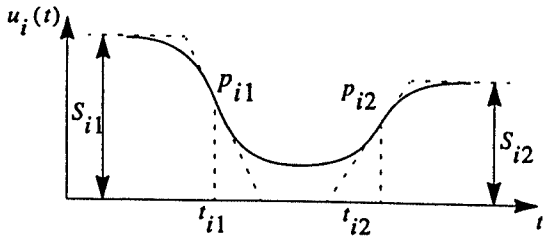


Figure 2 : Modélisation de séquences V_1CV_2.

Par contre comme nous l'avons déjà souligné, durant l'occlusion aucune information ne permet d'inférer directement la position de la partie stable (si elle existe) de l'occlusion. Il convient dès lors que celle-ci émerge de la procédure d'analyse et ne soit pas un paramètre de commande explicite.

3. ANALYSE

L'analyse consiste à déterminer les paramètres qui gouvernent les sigmoïdes. Il convient de considérer d'une part l'estimation des parties stables des sigmoïdes et d'autre part, des pentes et points d'inflexion.

L'estimation des parties stables des sigmoïdes (S_{i1} et S_{i2}) est réalisée par un réseau connexionniste entraîné pour réaliser l'inversion acoustico-articulatoire à partir des trois premiers formants (Jospa et al. [2], George et al. [1], Soquet [6]). Les avantages d'une telle méthode résident essentiellement dans la capacité du réseau d'apprendre des patrons caractéristiques de la transformation acoustico-articulatoire et par la suite à généraliser à bon escient cette connaissance explicite pour fournir, au départ de triplets de formants extraits de parole naturelle, des formes réalistes de conduit vocal.

Les parties stables ayant été estimées, il convient ensuite de déterminer les valeurs des pentes et des points d'inflexion.

Le schéma de principe de la méthode est présenté à la figure 3. Les pentes et points d'inflexion des sigmoïdes sont estimés par analyse-synthèse de manière à minimiser l'erreur sur les formants et ce, sur la transition complète.

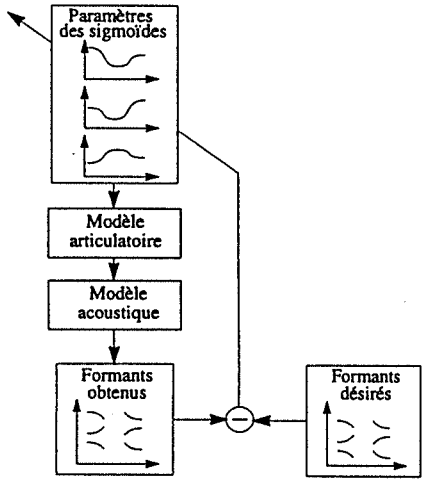


Figure 3 : Schéma de principe de l'estimation des pentes et des points d'inflexion par analyse-synthèse.

Ces paramètres permettent d'obtenir l'évolution temporelle des commandes articulatoires u(t) et des formants y(t). Ces derniers sont comparés aux formants désirés $y^d(t)$ extraits du signal de parole. Une erreur est calculée et un processus d'optimisation (par descente de gradient) est mis en place pour la minimiser.

4. RÉSULTATS ET DISCUSSION

Dans le cadre de cette étude, nous avons choisi comme modèle articulatoire le modèle à régions (Mrayati et al. [4]). Nous utiliserons le modèle à huit régions, soit :

$$u = [R_1, R_2, R_3, R_4, R_5, R_6, R_7, R_8] \quad (3)$$

où R_i représente la section de la $i^{ème}$ région. Les régions sont numérotées de la glotte (R_1) aux lèvres (R_8).

Nous avons enregistré un corpus constitué de logatomes V_1V_2 et V_1CV_2 répétés par un locuteur masculin où C représente une des six plosives [p, t, k, b, d, g] et V1,V2 une des voyelles orales [a, œ, o, y, u]. Une étude préliminaire des résultats obtenus sur ce corpus s'avère tout à fait encourageante.

La figure 4 montre l'évolution temporelle des huit paramètres de commande du modèle à régions obtenue pour une transition [a u]. La région R_1 a été fixée à 2.1 cm^2. On peut observer la bonne adéquation entre les formants mesurés sur le logatome et les formants obtenus par modélisation des trajectoires articulatoires. Au niveau articulatoire, (**i**) les parties stables des deux voyelles sont réalistes et comparables aux configurations décrites dans la littérature (Soquet [6]), et (**ii**) les transitions articulatoires permettent de rendre compte des transitions acoustiques mesurées. Les informations relatives aux pentes et aux points d'inflexion devraient nous permettre d'étudier les stratégies articulatoires utilisées par différents locuteurs pour réaliser des transitions vocaliques.

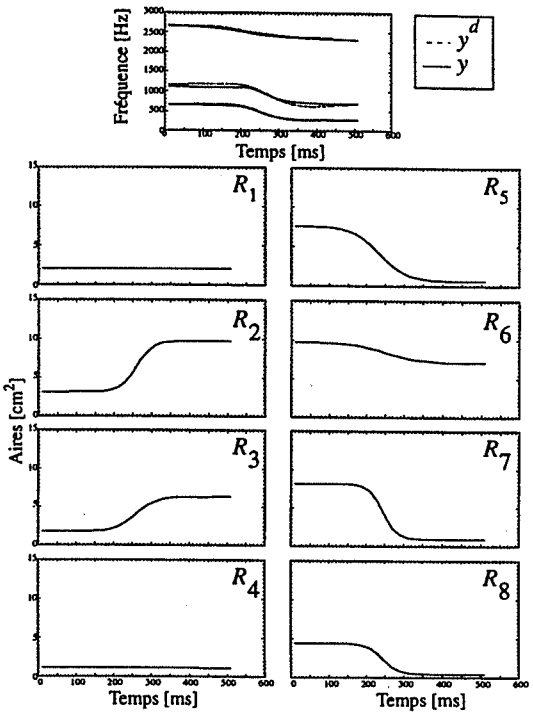


Figure 4 : Evolution temporelle des 8 régions du modèle à régions obtenues par analyse pour une transition [a u].

La figure 5 montre l'évolution temporelle des huit paramètres de commande du modèle à régions obtenue pour une transition [ato]. La zone centrale (entre pointillés) indique la plage temporelle durant laquelle les formants ne peuvent être mesurés. Cet intervalle de temps correspond approximativement à la durée de l'occlusion du conduit vocal. La région R_1 a été fixée à 2.1 cm^2.

Les trajectoires acoustiques et articulatoires sont tout à fait satisfaisantes. L'occlusion est complète pour la région R_6, ce qui est caractéristique du lieu d'articulation des occlusives dentales.

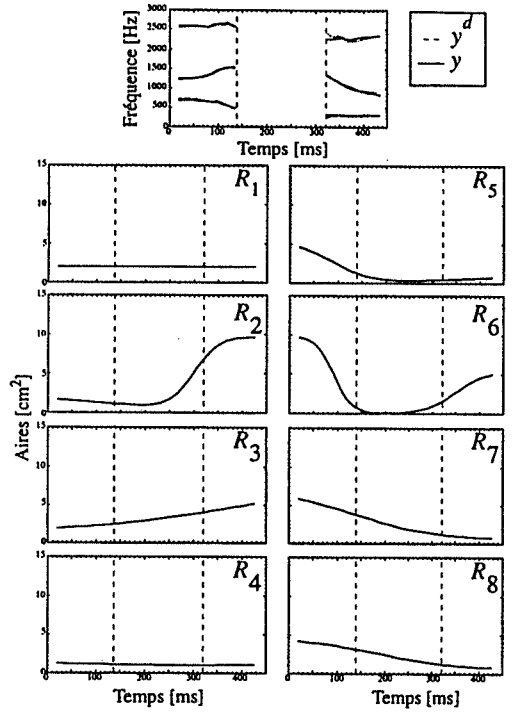


Figure 5 : Evolution temporelle des 8 régions du modèle à régions obtenue par analyse pour une transition [ato].

5. CONCLUSION

Dans ce travail, nous avons proposé une méthode de modélisation et d'analyse de séquences V_1V_2 et V_1CV_2. La modélisation consiste à décrire ces séquences en termes de l'évolution dans le temps des paramètres de commande d'un modèle de conduit vocal. Chaque paramètre de commande est exprimé à l'aide d'une combinaison de sigmoïdes. L'analyse permet, à partir du signal de parole, de déterminer les valeurs des paramètres de ces sigmoïdes. Nous avons choisi comme indices acoustiques les fréquences des trois premiers formants et comme paramètres articulatoires les sections du modèle à régions. Cette méthode a été appliquée avec succès à l'analyse de séquences V_1V_2 et V_1CV_2.

Les résultats préliminaires montrent la bonne adéquation entre les formants mesurés sur le logatome et les formants obtenus par modélisation des trajectoires articulatoires au moyen des modèles proposés. Au niveau articulatoire, **(i)** les parties stables des voyelles sont réalistes, **(ii)** les transitions articulatoires permettent de rendre compte des transitions acoustiques mesurées, et **(iii)** l'extrapolation des trajectoires articulatoires durant l'occlusion se montre prometteuse à la fois pour étudier les mécanismes de la production des occlusives, mais aussi, pour tenter de déterminer le lieu d'articulation des occlusives.

6. BIBLIOGRAPHIE

[1] M. George, P. Jospa, et A. Soquet, "Estimation de trajectoires articulatoires à l'aide d'un réseau de neurones," Actes des 20èmes Journées d'études sur la parole, Trégastel, pages 427-432, 1994.

[2] P. Jospa, A. Soquet, and M. Saerens, "Acoustical sensitivity functions and the control of the vocal tract model," Signal Processing VI: Theories and Applications, J. Vanderwalle, R. Boite, M. Moonen, A. Oosterlinck (editors), Morgan Kaufmann Publishers, pages 319-327, 1992.

[3] T. Kaburagi, et M. Honda, "Determination of sagittal tongue shape from positions of points on the tongue surface," J. Acoust. Soc. Am., vol. 96, n°3, pages 1356-1366, 1994.

[4] M. Mrayati, R. Carré, et B. Guérin, "Distinctive regions and modes: a new theory of speech production," Speech Communication, vol. 7, pages 257-286, 1988.

[5] W. Serniclaes, "Etude expérimentale de la perception de trait de voisement des occlusives du Français," Thèse de Doctorat, Université Libre de Bruxelles, 1987.

[6] A. Soquet, "Etude comparée de représentations acoustiques et articulatoires du signal de parole pour le décodage acoustico-phonétique : application à la classification de voyelles et à la détermination du lieu d'articulation des occlusives," Thèse de Doctorat, Université Libre de Bruxelles, 1995.

INFLUENCE DE LA VITESSE D'ÉLOCUTION ET DE L'ACCENT SUR DES CIBLES VOCALIQUES ESTIMÉES AUX NIVEAUX ACOUSTIQUE ET QUASI ARTICULATOIRE

Michel PITERMANN, Sorin CIOCEA, Jean SCHOENTGEN*

Université Libre de Bruxelles, Institut des Langues Vivantes et de Phonétique, CP 110 – 50, Av. F.D. Roosevelt – B-1050 Bruxelles – BELGIUM – e-mail : mpiter@ulb.ac.be

ABSTRACT

One of the authors recently studied the influence of speaking rate and stress pattern on formant frequencies and the corresponding target estimates of vocoïds [ɑ] and [ɛ] in an [i_i] context [Pitermann 1996]. Formant targets were estimated by means of the long-term behaviour of formant transition models. The results show that formant frequencies and target estimates vary with speaking rate and stress pattern. In the present article, we analyze the quasi articulatory targets estimated for the same corpus. The quasi articulatory trajectories were analytically calculated by means of acoustic-articulatory inversion carried out on a vocal-tract area-function model. Targets were estimated by means of asymptotes or steady-state solutions of transition models fitted to the trajectories of the palatal region. The results suggest that the variability, with respect to speaking rate and stress pattern, of the targets of the trajectories of the palatal region cross-section is lower than the variability of the extrema of the trajectories themselves.

1. INTRODUCTION

Récemment, un des auteurs [Pitermann 1996] a étudié l'influence de la vitesse d'élocution et de l'accent d'insistance sur les fréquences, et les cibles, des premier et deuxième formants des vocoïdes [ɑ] et [ɛ] dans un contexte [i_i] [Lindblom 1963; Gay 1981; Lindblom 1983]. Les cibles des formants ont été estimées en modélisant la transition entre le premier [i] et le [ɑ] ou [ɛ] suivant. Elles ont été assimilées au comportement à long terme du formant dans le cadre de quatre modèles de transition différents. Les résultats ont montré que: (a) les fréquences des deux premiers formants et les estimations des cibles dépendaient de la vitesse et de l'accent. Pour certaines combinaisons de débit et d'accent, la variabilité des cibles était moins importante que celle des valeurs brutes des fréquences, mais une analyse statistique n'a pas permis de conclure sans équivoque à l'existence de cibles invariantes par rapport à l'accent ou la vitesse; (b) les quatre modèles de transition donnaient lieu à des résultats semblables. Ces résultats suggèrent que les segments vocaliques ne sont pas caractérisés par des cibles formantiques invariantes ou que les méthodes d'estimation des cibles étaient inadéquates. Dans cet article, nous continuons à explorer ces questions au niveau quasi articulatoire. En effet, nous avons calculé analytiquement les aires des sections d'un modèle du conduit vocal de manière à faire correspondre ses fréquences propres aux fréquences des trois premiers formants observées. Ensuite, nous avons ajusté aux trajectoires de la section de la région palatale les quatre modèles de transition susmentionnés, et nous avons comparé les cibles quasi articulatoires ainsi estimées aux extrémums des trajectoires, extrémums localisés au milieu des segments vocaliques [ɑ] ou [ɛ]. Les résultats montrent que, pour un sous-groupe de modèles, les estimations de la cible étaient moins sensibles aux changements de débit ou d'accent que les extrémums des trajectoires.

2. MÉTHODE

2.1. Corpus

Deux locuteurs francophones ont produit les vocoïdes [ɑ] et [ɛ] dans un contexte [i_i] inséré dans une phrase porteuse prononcée avec ou sans accent d'insistance sur le vocoïde et à différentes vitesses d'élocution contrôlées par mé-

*Fonds National de la Recherche Scientifique, Belgique

tronome (locuteur A : 10 vitesses, locuteur B : 9 vitesses). Chaque combinaison {accent, débit} a été répétée 30 fois au moins. Ensuite, les fréquences des formants ont été estimées à l'aide d'une analyse par prédiction linéaire. Les stimuli pour lesquels les instructions données aux locuteurs et l'identité perçue du vocoïde (déterminée par un panel de sept juges) ne concordaient pas ont été repérés dans le corpus et écartés de la discussion.

2.2. Inversion acoustico-articulatoire

L'objectif de l'inversion acoustico-articulatoire était ici de calculer les valeurs des sections d'un modèle du conduit vocal de manière à ce que ses trois premières fréquences propres soient identiques aux trois premiers formants observés. Le modèle était du type Kelly-Lochbaum à six tubes uniformes enchaînés de longueurs égales et de sections variables. La méthode de transformation fréquence-aire que nous avons proposée est directe et ne fait appel ni à une estimation des fonctions de sensibilité, ni à un optimiseur, ni à une fonction coût ad hoc [Schoentgen and Ciocea 1995a; Schoentgen and Ciocea 1995b]. Son point de départ est la condition des fréquences propres $F(S_i, l_i, f_j) = 0$ obtenue en multipliant les matrices de transfert des tubes individuels et en imposant des conditions aux limites à la glotte et aux lèvres. La condition $F(S_i, l_i, f_j) = 0$ lie les longueurs l_i et les sections S_i des tubes concaténés aux fréquences propres f_j. Dans notre cas, les longueurs l_i sont une fraction fixe de la longueur totale qui dépend de la section aux lèvres. Un système d'équations algébriques reliant à un instant donné les dérivées par rapport au temps des fréquences des formants f_j et des paramètres de la fonction d'aire $X_i = \{l_i, S_i\}$, peut être obtenu en appliquant à la condition $F = 0$ un théorème dit « règle de chaînage ». Au sein de ce système d'équations linéaires, les seules inconnues sont les dérivées $\frac{dX_i}{dt}$. La solution du système d'équations est obtenue à l'aide de la méthode de décomposition en valeurs singulières qui, en général, fournit une infinité de solutions parmi lesquelles une solution unique est choisie moyennant des contraintes additionnelles. Les valeurs des paramètres X_i sont alors calculées numériquement à partir des dérivées $\frac{dX_i}{dt}$ et les opérations recommencent avec le calcul, à un instant ultérieur, des dérivées par rapport au temps de la condition F.

TAB. 1 - *Équations des modèles de transition [équation aux différences linéaire (1) et non linéaire (2), fonction exponentielle (3) et logistique (4)] et des estimateurs de cible. Le symbole x_n représente la valeur d'un formant ou d'une section à l'instant n ; a_i, a, b, c, d, p, q et t_0 sont les paramètres des modèles; k est l'ordre de l'équation aux différences linéaire (1); Δt est le pas d'échantillonnage (9 ms); x^* est l'estimation de la cible.*

	modèles de transition	estim. de cibles
(1)	$x_n = a_0 + \sum_{i=1}^{k} a_i x_{n-i}$	$x^* = \frac{a_0}{1-\sum_{i=1}^{k} a_i}$
(2)	$x_n = c + \frac{a(x_{n-1}-c)}{ad+(1-d)(x_{n-1}-c)}$	$x^* = c + a$
(3)	$x_n = x^* + pe^{-qn\Delta t}$	$x^* = x^*$
(4)	$x_n = c + \frac{a}{1+e^{-b(n\Delta t - t_0)}}$	$x^* = c + a$

2.3. Estimation des cibles

Les quatre modèles de transition étaient les suivants: une équation aux différences linéaire (1), une équation aux différences non linéaire (2), une courbe exponentielle (3), et une courbe logistique (4), solution de l'équation non linéaire (2) (cf. Tableau 1). Les transitions ajustées étaient les trajectoires de la fréquence du premier formant F_1 ou de la section de la région palatale du modèle de Kelly-Lochbaum, c.-à-d. la section du tube numéro 5 en comptant à partir de la glotte. Les modèles du type logistique ont été ajustés aux transitions complètes et ceux du type exponentiel aux demi-transitions finales entre [i] et [ɑ] ou [ɛ]. Chaque ajustement reposait sur l'ensemble des 30 répétitions de chaque stimulus afin d'augmenter la fiabilité des estimations. Les cibles étaient assimilées aux asymptotes des courbes exponentielle et logistique et aux solutions stationnaires des équations linéaire et non linéaire. Les indices utilisés pour caractériser les vocoïdes [ɑ] ou [ɛ] étaient le maximum de la trajectoire de la fréquence du premier formant F_1 et le maximum de la section « palatale ».

3. RÉSULTATS ET DISCUSSIONS

En ce qui concerne la variabilité des cibles quasi articulatoires par rapport au débit de parole et à l'accent d'insistance, les meilleurs résultats ont été obtenus grâce à l'équation aux différences non linéaire (2) appliquée à la section de la région palatale du modèle de Kelly-Lochbaum. Nous limiterons donc la présentation des résultats et leur discussion à ce cas.

La Figure 1 montre, en fonction du débit

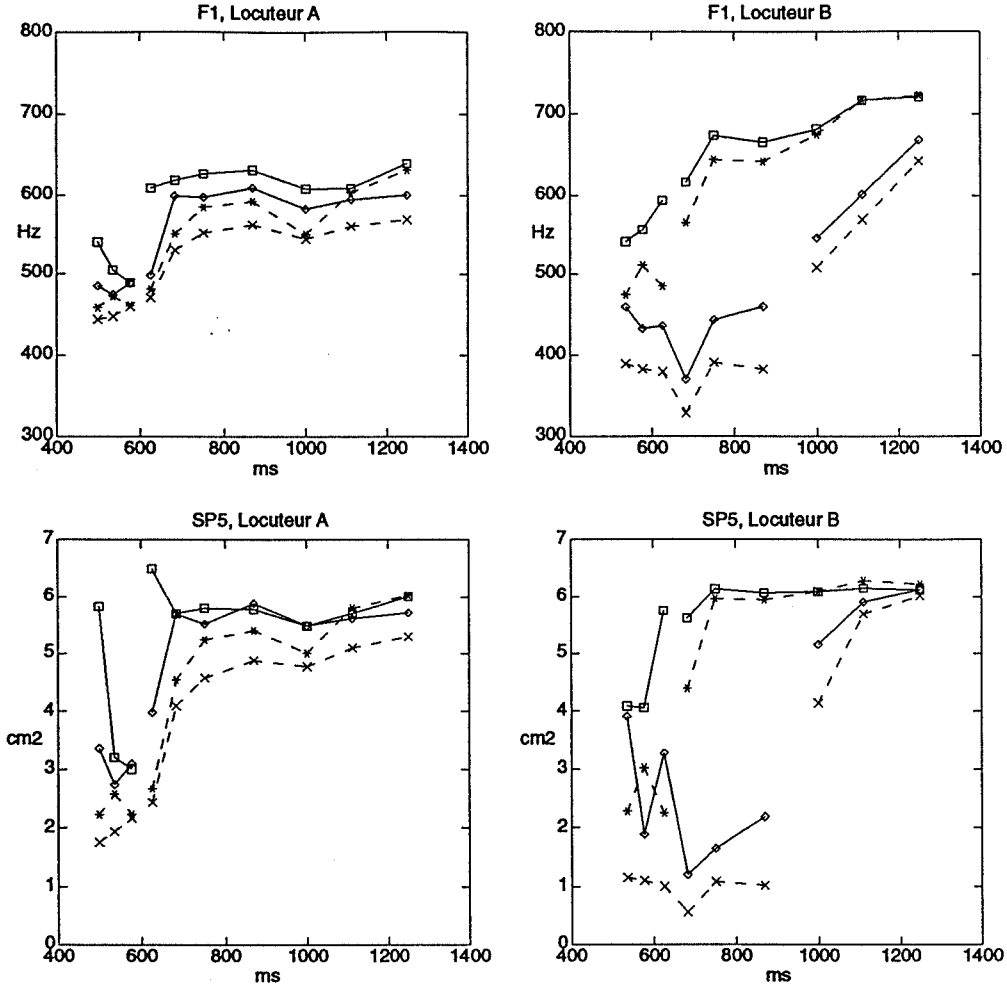


FIG. 1 - *Comparaison entre la fréquence du premier formant (F_1), la section « palatale» (SP_5) et leurs cibles pour le vocoïde [ɑ] en fonction de la vitesse d'élocution ('∗': F_1 ou SP_5 pour le [ɑ] accentué, '×': F_1 ou SP_5 pour le [ɑ] non accentué, '□': cible pour le [ɑ] accentué, et '◇': cible pour le [ɑ] non accentué). L'interruption des graphes sépare les stimuli correctement et incorrectement perçus. Les cibles sont reliées par des traits pleins et les observations par des traits discontinus.*

d'élocution, pour le vocoïde [ɑ], les grandeurs suivantes: (a) Les fréquences de F_1 (∗ vocoïde accentué; × vocoïde non accentué); (b) les cibles estimées des fréquences de F_1 (□ accentué; ◇ non accentué); (c) les sections du tube numéro 5 du modèle de Kelly-Lochbaum (∗ accentué; × non accentué); (d) les cibles estimées des trajectoires de la section du même tube (□, accentué; ◇, non accentué). Dans la Figure 1, le débit est indiqué en millisecondes entre deux battements de métronome. Les débits lents sont à droite des diagrammes. Le choix de présenter conjointement les changements du premier formant et de la section de la région palatale est conforme à la règle selon laquelle la fréquence du premier formant reflète, en première approximation, l'ouverture de la moitié antérieure du conduit vocal.

Pour [ɛ], les extrémums et cibles des trajectoires étaient généralement confondus, et leurs intervalles de variation étaient disjoints de ceux des [ɑ] correctement perçus (résultats non présentés).

Comparons d'abord, pour le vocoïde [ɑ], l'influence de l'accent sur les grandeurs suivantes.

a) la fréquence de F_1;

b) l'estimation de la cible de F_1;

c) la section de la région palatale du modèle (tube numéro 5);

d) l'estimation de la cible de la section.

Aussi, limitons la discussion aux classes des [ɑ] correctement perçues, c.-à-d. aux sept débits les plus lents pour le locuteur A et aux six (cas

accentué) et trois (cas non accentué) débits les plus lents pour le locuteur B. Pour le locuteur A, les valeurs médianes des différences entre les cas accentué et non accentué étaient les suivantes.

a) 30 Hz pour les fréquences de F_1 ;

b) 25 Hz pour les estimations des cibles de F_1 ;

c) 0.53 cm^2 pour la section « palatale » ;

d) 0.09 cm^2 pour les cibles estimées de la section « palatale ».

Pour le locuteur B, les valeurs médianes des mêmes différences étaient respectivement : a) 147 Hz, b) 115 Hz, c) 0.59 cm^2 et d) 0.24 cm^2.

Comparons ensuite, toujours pour le vocoïde [ɑ], l'influence du débit d'élocution sur les mêmes grandeurs, c.-à-d. F_1, cible de F_1, section « palatale » et cible de la section « palatale ». D'abord, la Figure 1 montre que la fréquence de F_1 et la section « palatale » dépendaient du débit, et que les cibles de F_1 étaient moins influencées par le débit que la fréquence de F_1, observations confirmées par une analyse de la variance ou de régression [Pitermann 1996]. Ensuite, la Figure 1 suggère que l'influence du débit sur la cible de la section « palatale » était aussi réduite. Le graphe, de la dépendance des cibles ou des sections par rapport au débit, apparaît horizontal et, surtout, cette horizontalité est interrompue à partir des débits des stimuli incorrectement perçus. Au-delà de ce point de rupture, la relation entre débit et cibles devient erratique.

Ainsi, la réduction de la variabilité des cibles estimées des trajectoires des formants trouve son pendant au niveau des trajectoires des indices quasi articulatoires. Ce résultat n'était pas assuré a priori. En effet, les non linéarités des liens entre formants et forme du conduit auraient pu amplifier, au lieu de réduire, la variabilité des trajectoires ou cibles. La bonne tenue des indices quasi articulatoires est confirmée par le fait que, parmi les quatre modèles de transition, les performances de modélisation du groupe des non linéaires (équation (2) et courbe logistique) étaient meilleures que celles du groupe des linéaires (équation (1) et courbe exponentielle). Cette différenciation des performances n'a pas été observée lorsque les modèles étaient appliqués aux indices acoustiques, quoique les deux groupes de modèles décrivent des dynamiques qualitativement différentes.

4. CONCLUSION

Les résultats présentés sont préliminaires. Ils devront être confrontés à des mesures faites sur d'autres modèles du conduit vocal et, surtout, à des données articulatoires obtenues par mesure directe. Provisoirement, tenant compte des données acoustiques et quasi articulatoires dont nous disposons, nous ne pouvons pas conclure à l'existence de cibles vocaliques invariantes par rapport à l'accent d'insistance ou au débit d'élocution. Néanmoins, les résultats suggèrent que des zones de variabilité réduite existent pour certaines cibles, la réduction de la variabilité étant liée à l'extrapolation de la trajectoire des indices acoustiques ou articulatoires au voisinage de cibles non atteintes.

BIBLIOGRAPHIE

Gay, T. (1981). Mechanisms in the control of speech rate. *Phonetica* **38**, 148–158.

Lindblom, B. (1983). *Economy of Speech Gestures.* P. F. MacNeilage, Springer-Verlag. New York.

Lindblom, B. E. F. (1963). Spectrographic study of vowel reduction. *The Journal of the Acoustical Society of America* **35**(11), 1773–1781.

Pitermann, M. (1996). *Évaluation expérimentale de la théorie des cibles formantiques dans le cadre de la production des voyelles.* Ph. D. thesis, Université Libre de Bruxelles.

Schoentgen, J. and S. Ciocea (1995a). Direct calculation of the vocal tract area function from measured formant frequencies. In *Eurospeech'95 Proceedings*, Volume 1, pp. 745–748.

Schoentgen, J. and S. Ciocea (1995b). Kinematic acoustic-to-geometric mapping. In *Proceedings of the XIIIth International Congress of Phonetic Sciences*, Volume 2, pp. 194–197.

INTÉRÊT DE L'IMAGERIE PAR RÉSONANCE MAGNÉTIQUE DANS L'EXPLICATION PHYSIOLOGIQUE DU FORMANT DU CHANTEUR

Claire Pillot

Université de Paris III - Institut de Phonétique de Paris - 19, rue des Bernardins - 75005 PARIS

Tel.: 43 26 37 80 - Fax: 43 29 70 13 - e-mail: pillot@msh-paris.fr

ABSTRACT

Magnetic Resonance Imaging (MRI) slices of french vowels [a], [i] and [o] emitted by a professional bass singer show a great number of articulatory modifications from speech to singing, which interact between each other. The transfer function estimated from physiological data (using the Maeda's vocal tract acoustic simulation program) shows several acoustic modifications. The most significative is the bringing together of F3 and F4. This phenomenon corresponds to the well known singing-formant.

MOTS-CLÉS

Production de parole - phonétique - chant - Imagerie par Résonance Magnétique - Simulation acoustique - Formant du chanteur.

INTRODUCTION

Inventée par Bloch et Purcell en 1946, la "Résonance Magnétique Nucléaire" (RMN) est utilisée pour la première fois à Tokyo en 1986 en vue d'explorer le conduit vocal (Sulter, 1992; Baer, 1987). L'objet de cette étude est de mettre en évidence, au moyen de l'Imagerie par Résonance Magnétique (IRM), la nature des modifications articulatoires d'une voyelle parlée à son équivalente chantée chez un chanteur lyrique professionnel, d'établir les fonctions d'aire et de transfert issues des données physiologiques et de corréler certains paramètres articulatoires et acoustiques.

MÉTHODE

Les raisons du choix de l'IRM comme procédé préférentiel d'exploration du conduit vocal tiennent à ses nombreux avantages par rapport à la radiographie classique: absence de radiations ionisantes, meilleure définition des parties molles et multiplicité des plans de coupe.

Le sujet choisi est une basse professionnelle dépourvu de contre-indications (port d'objets ferro magnétiques, claustrophobie). En raison de grandes variabilités morphologiques d'un sujet à l'autre, nous avons préféré étudier les productions répétées d'un seul sujet plutôt que de comparer les émissions uniques de plusieurs chanteurs. Notre expérience peut donc s'apparenter à une étude de cas en raison de l'ampleur du travail d'analyse, du coût de l'examen, de la disponibilité et du nombre limité de machines, de ses conditions contraignantes, et du faible nombre de sujets dépourvus de contre-indications acceptant ce type d'investigations.

Le sujet, allongé, a émis deux fois les voyelles françaises [a], [i] et [o] à 100Hz (et 150Hz en plus pour le [a]) dans un placement vocal parlé d'une part, et chanté d'autre part. On peut s'interroger sur l'influence de la position allongée du sujet sur les productions articulatoires et sonores obtenues. Ne sont-elles pas ainsi éloignées de leur contexte naturel d'émission ? Detweiler (1994) affirme que l'analyse des productions chantées émises en position allongée vérifie la présence d'un formant du chanteur stable, statistiquement non distinct de celui qui est obtenu en position debout. La position allongée est donc supposée ne pas exercer d'influence notable sur les résultats obtenus. Un chanteur d'Opéra se doit d'ailleurs de chanter dans toutes les positions sur scène ...

Les coupes sagittales médianes correspondantes ont ainsi été acquises (Champ magnétique: 0,5T d'une machine Gyrex V; ET: 15ms; RT: 33ms; antenne de tête; matrice d'acquisition de 138 x 200 voxel; temps d'acquisition: 8s). Leurs contours ont permis de quantifier les modifications articulatoires. Les fonctions d'aire et de transfert ont ensuite

été obtenues grâce à l'utilisation du programme de simulation acoustique du conduit vocal de S. Maeda (Maeda, 1992).

RÉSULTATS

Analyse des images obtenues (données physiologiques)

Il existe un grand nombre de modifications articulatoires d'un [o] parlé au [o] chanté (figure 1). Ces changements se retrouvent également dans les voyelles [a] et [i]. ...

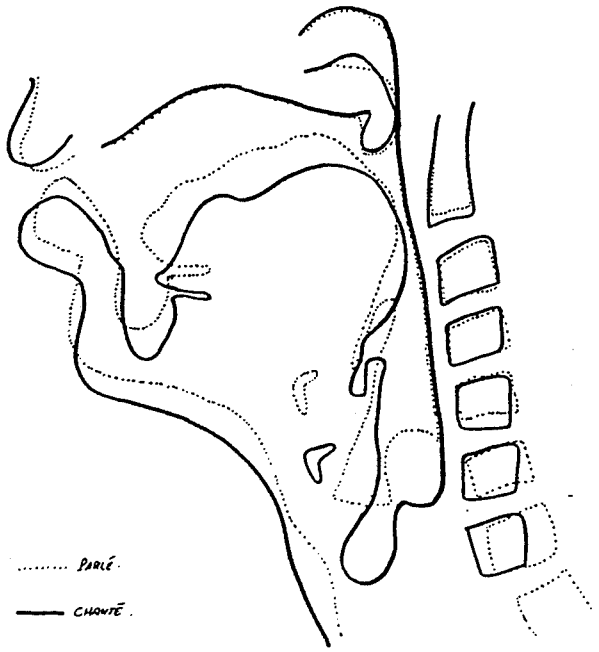


Figure 1: Superposition des contours issus des coupes sagittales médianes d'IRM d'un [o] parlé et chanté à la même fréquence (100 Hz) montrant l'ensemble des différences articulatoires obtenues.

Le larynx et l'os hyoïde s'abaissent significativement de 2 cm pour l'ensemble des productions (t_7=-10,7; p=0,0001) et le tube laryngé s'allonge de 0,5 cm en moyenne. Le corps de la langue s'abaisse généralement de 0,6 cm de la parole au chant (t_7=4,76; p=0,002) mais cette tendance est moindre pour le [i]. Le point de constriction lingual se postériorise et la mandibule s'abaisse généralement de 0,6 cm (t_7=-3,8; p=0,006), excepté pour le [i]. Les lèvres sont plus ouvertes et plus avancées. L'orifice ary-épiglottique est plus étroit.

De plus, le chant entraîne une réduction de la courbure cervicale pour l'ensemble des voyelles émises, ainsi qu'un voile du palais plus fortement collé contre la paroi pharyngée postérieure. Toutes nos données se sont révélées reproductibles et on observe peu de modifications en fonction de la hauteur du son pour un même placement donné ("parlé" ou chanté).

On observe en outre des corrélations statistiquement significatives entre l'abaissement laryngé et: celui de l'os hyoïde (r_8=0,9), la longueur du larynx (r_8=-0,8), la protrusion labiale (r_8=-0,68) et l'ouverture mandibulaire (r_8=0,88). De plus, plus la mandibule est ouverte, plus le dos de la langue s'abaisse et s'accompagne d'une postériorisation du point de constriction lingual. Les modifications articulatoires sont de plus grande ampleur pour [a], et légèrement moindres pour [i].

Analyse des fonctions d'aire

Celle-ci fait apparaître une augmentation de la longueur du conduit vocal de 2 cm en moyenne de la parole au chant, résultant principalement en un accroissement de la cavité postérieure (t_7=-2,6; p<0,04). On observe en outre une baisse du rapport entre les sections laryngée et pharyngée dans nos données. Celles-ci sont reproductibles et ne subissent pas d'influence de la hauteur du son.

Analyse des fonctions de transfert (données acoustiques)

La figure 2 montre, de la parole au chant: une baisse significative des quatrième (t_5=3,6; p=0,01) et sixième formants (t=7,05; p=0,01), une réduction des distances fréquencielles entre F3/F4 d'une part, F5/F6 d'autre part (sauf pour [i]), et l'absence d'apparition d'un formant supplémentaire comme le suggère la littérature (Sundberg, 1987). Ces tendances sont reproductibles et ne subissent pas d'influence de la hauteur du son.

Corrélation entre les paramètres articulatoires et acoustiques

Nous notons une relation importante entre l'abaissement de F4 et celui du larynx (r_{20}=0,88). Il en est de même pour F6 (r_{12}=0,86).

En outre, la diminution du rapport entre les sections laryngée et pharyngée signifie une augmentation de la discontinuité entre les cavités laryngée et postérieure; on note qu'une augmentation de cette discontinuité est en lien avec un rapprochement notable des troisième et quatrième formants (r_{16}=0,76).

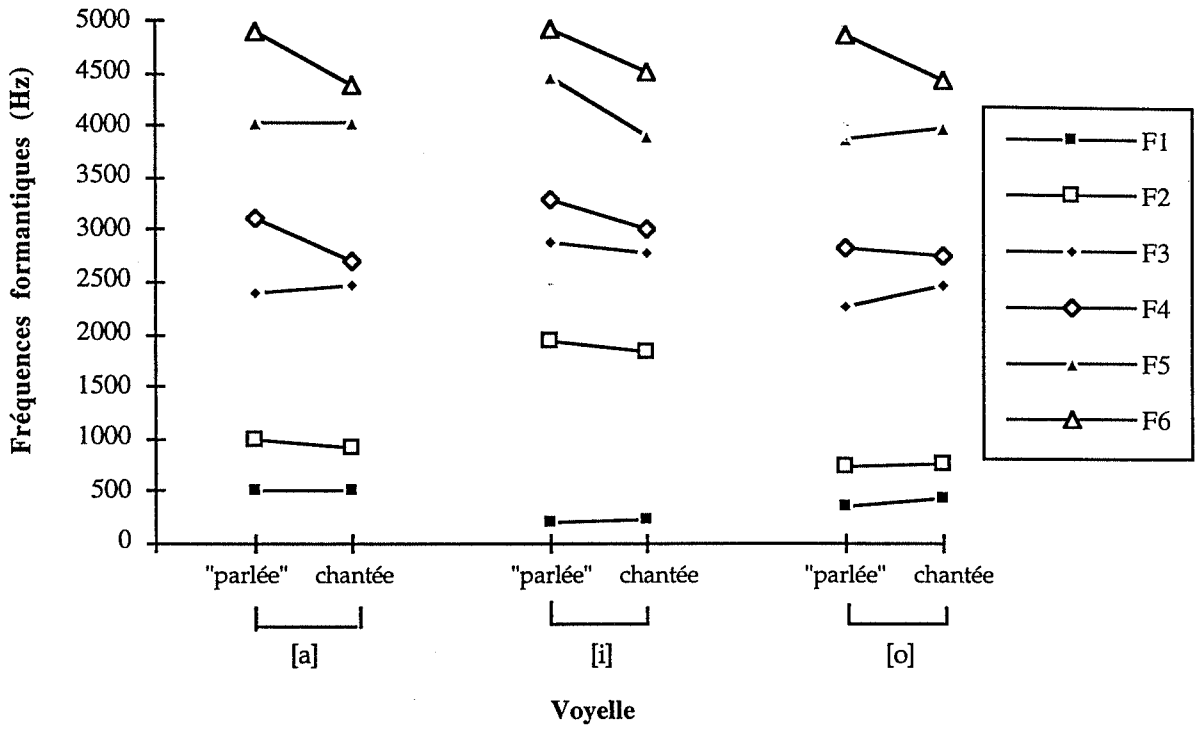


Figure 2: Effet du chant sur la fréquence des formants des trois voyelles simulées (données moyennes).

DISCUSSION

Le rapprochement de F3 et F4 entre 2500Hz et 3000Hz semble être la modification acoustique la plus significative que nous ayions observée de la parole au chant. Or, on retrouve un tel accolement sous forme d'un formant unique dans la même zone de fréquences, dans les voyelles et les autres productions vocales enregistrées de notre sujet (figure 3): la comparaison des voyelles chantées enregistrées et simulées pour la basse professionnelle montre des fréquences formantiques quasiment identiques. De plus, les voyelles parlées simulées et enregistrées révèlent des spectres semblables, et notamment une absence de rapprochement entre F3 et F4.

La littérature connaît depuis longtemps ce phénomène acoustique intitulé "formant du chanteur" (Sundberg, 1987). Ce renforcement énergétique (vers 2500 Hz pour les voix les plus graves et 3200 Hz pour les plus aiguës) ne varie pas quelle que soit la voyelle chantée, mais n'apparaît en général que chez les sujets possédant une voix travaillée selon la technique du chant lyrique classique.

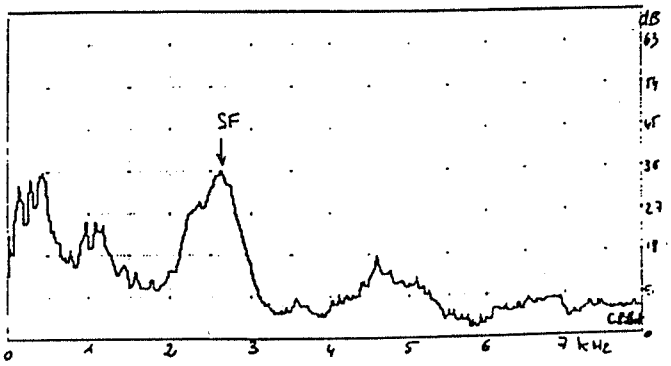


Figure 3: Spectre à long terme moyenné sur 18s d'un morceau chanté par notre sujet, montrant l'émergence d'un formant unique vers 2500 Hz.

Le formant du chanteur se situe à une zone où la sensibilité auditive est maximale, ce qui laisse penser qu'il joue un rôle perceptif important, permettant au chanteur de mieux se faire entendre avec un accompagnement orchestral, mais donnant également à sa voix une sonorité et une couleur particulières (Pillot, 1995).

Quels sont les principaux facteurs permettant la production du formant du

chanteur? Notons que la source laryngée produit plus d'énergie vers 3 KHz chez les sujets possédant le formant du chanteur: Si l'on considère l'influence des paramètres de l'onde de débit sur les formants supérieurs, on constate une diminution du quotient d'ouverture (rapport des temps d'ouverture et de fermeture sur la durée totale du cycle vibratoire cordal) et une augmentation du quotient de dissymétrie (rapport du temps d'ouverture sur le temps de fermeture des cordes vocales) permettant aux spectres des sons résultants de posséder des harmoniques aigus plus intenses. (Pillot, 1995). Ces phénomènes seuls ne permettent cependant pas la création d'un pic tel que le formant du chanteur: celui-ci est aussi un problème d'articulation.

Il semble que l'abaissement laryngé soit la principale cause de sa génération. Pour Sundberg (1972), cet abaissement est associé à un élargissement des sinus piriformes et des ventricules de Morgagni, ce qui aboutit selon lui à un rapprochement des troisième, quatrième et cinquième formants. Mais l'IRM nous révèle ici en détail des ajustements compensatoires inattendus du conduit vocal qui accompagnent l'abaissement laryngé: nos données montrent que les modifications articulatoires permettant au formant du chanteur de se former ne sont pas liées uniquement au larynx mais au conduit vocal dans son ensemble.

La simulation des voyelles à partir des contours articulatoires obtenus montre que le formant du chanteur ne résulte, chez notre sujet, que du rapprochement de F3 et F4 pour les voyelles françaises [i], [a] et [o] sans que l'existence d'un formant supplémentaire (Sundberg, 1987) ne soit mis en évidence. On objective en outre le rapprochement de F5 et F6 (excepté pour la voyelle [i]) créant un pic semblable au formant du chanteur à son octave, comme l'avait déjà remarqué Bartholomew (1934).

CONCLUSION

A l'heure actuelle, l'IRM demeure encore un moyen d'exploration du conduit vocal lourd de contraintes: impossibilité d'enregistrer les sons obtenus durant l'acquisition des images, longueur de l'examen, et importante durée des temps d'acquisition. L'impossibilité d'effectuer des études dynamiques de la forme du conduit vocal devrait être prochainement résolue grâce au développement de nouveaux matériels de traitement de l'image IRM.

Cependant, l'exploration des voyelles chantées au moyen de l'Imagerie par Résonance Magnétique nous a permis d'augmenter la découverte du nombre de possibilités articulatoires offertes par le conduit vocal, mais aussi de corréler ces indispensables données initiales à des caractéristiques acoustiques. Abondamment décrit par la littérature, le "formant du chanteur", constitué du rapprochement fréquenciel de F3 et F4, paraît donc être en relation avec toutes les modifications articulatoires décrites chez notre sujet. Sa création est également favorisée par des modifications de la source (Pillot, 1995). Des études semblables chez des chanteurs et chanteuses d'autres catégories et techniques vocales sont nécessaires afin de vérifier ces données.

REMERCIEMENTS

Cette étude a pu être réalisée grâce à la disponibilité de Jacqueline Vaissière et aux judicieux conseils méthodologiques du Professeur Philippe Halimi. Je remercie également pour leur aide Shinji Maeda, le Docteur Monique Elmaleh et le sujet qui a subi avec patience l'examen d'acquisition des images de Résonance Magnétique.

RÉFÉRENCES

SULTER A.M. & Collaborateurs (1992): "On the relation between the dimensions and resonance characteristics of the vocal tract: a study with MRI", *Magnetic Resonance Imaging*, vol 10, 365-373.

BAER T. et Coll (1987): Application of MRI to the analysis of speech production, *Magnetic Resonance Imaging*, vol 5, 1-7.

DETWEILER R.F. (1994): An Investigation of the laryngeal System as the Resonance Source of the Singer's formant, *Journal of voice*, vol 8, n°4, 303-313.

MAEDA S. (1992): "Modélisation articulatoire du conduit vocal", *Journal de Physique IV*, vol 2, 307-314.

SUNDBERG J. (1987): *The science of the singing voice*, Dekalb Illinois, Northern Illinois University Press, 216 p.

PILLOT C. (1995): Production and perception of the singing-formant, *ICPhs 95 Stockholm*, vol 1, 262-265.

SUNDBERG J. (1972): "Articulatory interpretation of the singing-formant", *JASA*, vol 55, n°4, 838-844.

BARTHOLOMEW W.T. (1934): A physical definition of "good voice quality" in the male voice, *JASA*, June 1934, vol VI, 25-33.

PILOTAGE DYNAMIQUE D'UN MODÈLE DE PRODUCTION

Laurence CANDILLE, Henri MÉLONI

Laboratoire d'informatique d'Avignon - 339, chemin des Meinajariès BP 1228 - 84140 AVIGNON Cedex 9

Tél.: 90 84 35 09 Fax: 90 85 34 01 - e-mail: candille@univ-avignon.fr

ABSTRACT

A number of experiments have shown that it is conceivable to use production models for speech recognition tasks [Rose, 1994; Candille, 1995]. We present here the first results of an adaptation of Maeda's statistic model to the constraints of the problem. It is particularly necessary to take into account the static and dynamic characteristics of the speaker.

1. INTRODUCTION

En vue de la reconnaissance automatique de la parole au moyen de modèles de production, nous présentons les premiers résultats d'une adaptation du modèle de Maeda (Maeda, 1979) aux contraintes du problème (prise en compte des caractéristiques statiques et dynamiques du locuteur). Le paramètre permettant d'utiliser des longueurs différentes du conduit vocal est ajusté pour chaque locuteur. Pour chacune des voyelles orales, une configuration optimale du modèle est définie de manière à minimiser la distance entre les paramètres acoustiques caractérisant les productions du locuteur et celles du modèle. Nous avons effectué des mesures sur les différents mouvements naturels des articulateurs (lèvres, langue et mâchoire) lors de la réalisation de diphones vocaliques. On constate fréquemment des non-linéarités dans les variations articulatoires et des désynchronisations entre les différents paramètres. Ces phénomènes justifient les différences entre les trajectoires réelles et celles produites par le modèle linéairement interpolé. La transposition, dans le modèle, des vitesses et des accélérations mesurées en divers points des articulateurs a permis de réaliser des transitions entre deux configurations cibles qui produisent des trajectoires acoustiques très proches de celles qui sont évaluées pour un locuteur.

2. ADAPTATION STATIQUE DU MODÈLE

Le modèle de Maeda est un modèle statistique élaboré à partir des radiographies d'un conduit vocal humain; il est donc nécessaire d'effectuer une adaptation de ses caractéristiques statiques afin d'accorder optimalement les configurations des sons à chaque locuteur (Payan, 1993).

2.1. Modification de la longueur du conduit

Pour chaque locuteur, le paramètre du modèle permettant de faire varier la longueur totale du conduit vocal est ajusté de manière à faire coïncider au mieux l'espace acoustique produit par le modèle avec celui du locuteur. Cet ajustement se fait en minimisant la distance entre les valeurs des trois premiers formants obtenues à partir de la configuration standard de la voyelle [y] et les valeurs calculées pour les réalisations de ce phonème par le sujet. Pour chaque formant, la distance entre la valeur proposée par le modèle et celle qui lui est associée dans le signal est égale au nombre de bandes critiques qui les séparent.

2.2. Choix des configurations optimales

Pour chacune des voyelles orales, une configuration optimale du modèle est définie de manière à minimiser la distance entre les paramètres acoustiques caractérisant les productions du locuteur et celles du modèle. Cette configuration adaptée au locuteur est choisie dans un ensemble de formes du conduit qui sont des variations limitées autour d'un prototype standard de chaque voyelle. La qualité des voyelles ainsi déterminées est validée d'une part, en vérifiant l'adéquation des fonctions d'aire (figure1) avec celles des configurations standards (vérification des trois points principaux : position et aire de la constriction, aire aux lèvres (Fant, 1960; Boë, 1992)) et, d'autre part, en s'assurant de la qualité synthétique des sons produits par le modèle. Par ailleurs, les configurations ont été choisies de telle sorte que les variations des commandes du modèle, lors du passage d'une configuration à l'autre, soient cohérentes avec les mouvements naturels des articulateurs du locuteur produisant la même séquence. La figure 2 présente comparativement dans les plans F1/F2 et F1/F3 les valeurs acoustiques des voyelles moyennes du

locuteur LC et celles obtenues avec le modèle (par rapport aux configurations standards et par rapport aux configurations adaptées).

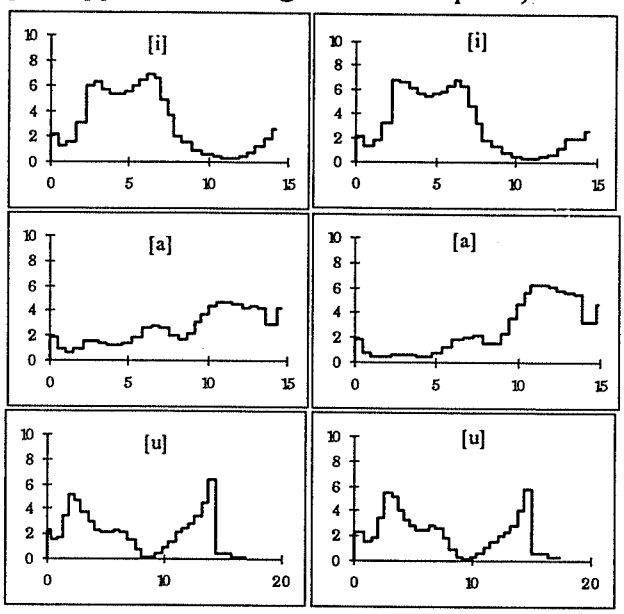


Figure 1. Comparaison des fonctions d'aires obtenues à l'aide du modèle de Maeda (à droite) adapté en longueur et celles déduites des configurations standards (à gauche) pour les voyelles [i], [a] et [u]. L'axe des abscisses représente la distance à la glotte en cm et l'axe des ordonnées l'aire des régions en cm^2.

Le cœfficient de longueur du modèle à été préalablement fixé pour toutes les configurations et les 10 prototypes standards sont extraits des 33 voyelles d'UPSID (Maddieson, 1986; Vallée, 1994).

Dans le plan F1/F2 les voyelles orales du français pour deux locuteurs (LC, locuteur féminin et TS, locuteur masculin) sont très bien représentées par les valeurs formantiques obtenues à partir des configurations adaptées. Dans le plan F1/F3 nous notons, pour les deux locuteurs, que le troisième formant des voyelles arrières [a], [o] et [ɔ] n'est pas atteint par les valeurs acoustiques des configurations adaptées du modèle.

Les résultats font apparaître que la somme des distances acoustiques entre les voyelles des configurations optimisées du modèle et les voyelles du locuteur est améliorée de plus de 70% par rapport à la même distance faisant intervenir les voyelles standards du modèle.

3. ADAPTATION DYNAMIQUE DU MODÈLE

Actuellement le modèle permet la représentation statique de toutes les voyelles du français mais ne propose pas une méthode pour le passage d'une forme du conduit vocal à une autre.

La production des transitions formantiques voyelle-voyelle par interpolation linéaire des

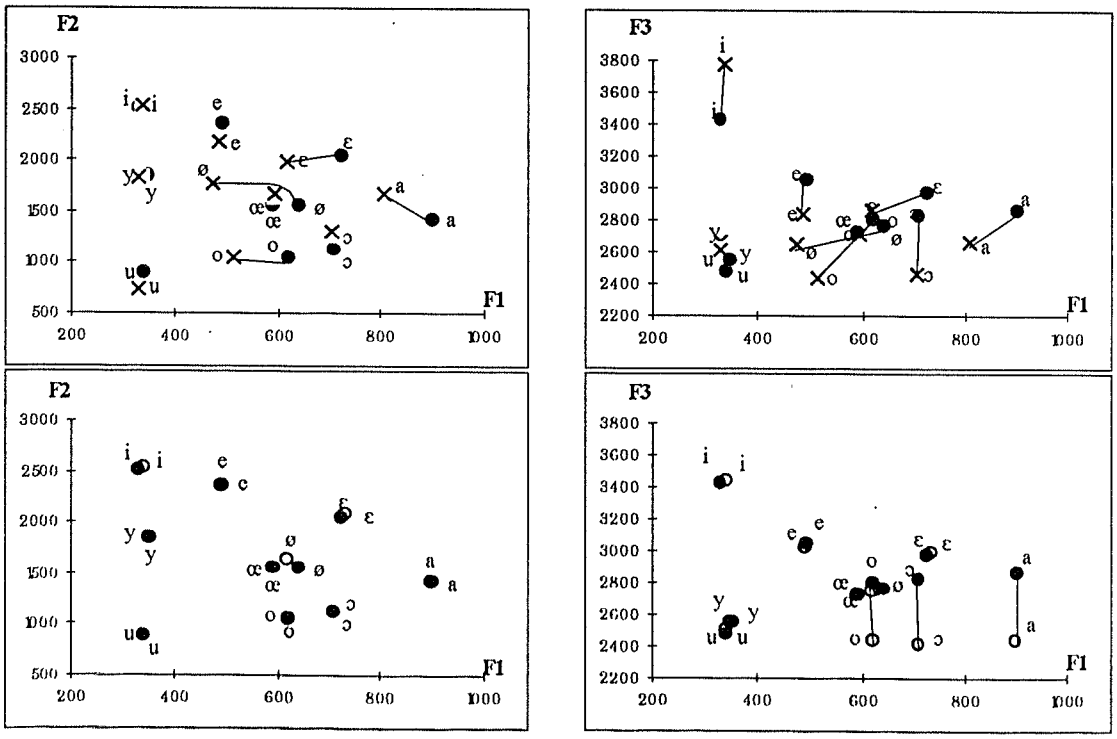


Figure 2 : Comparaison dans les plans F1/F2 (à gauche) et F1/F3 (à droite) des valeurs acoustiques des voyelles orales du français; celles du locuteur sont représentées par le caractère "•", celles des voyelles standards du modèle par "+" (figures du haut) et celles correspondant aux configurations adaptées par "o" (figures du bas).

paramètres entre les deux configurations cibles du modèle ne permet pas, dans certains cas, de reproduire précisément les trajectoires observées pour un locuteur (figure 4). Nous avons également étudié l'effet acoustique dans le plan F1/F2 produit par une variation de la vitesse de déplacement de chaque articulateur. Les résultats sont très nettement améliorés lorsque les articulateurs se déplacent de façon plus conforme à la réalité. La transition d'une voyelle à l'autre dépend donc de la coordination et de la vitesse de déplacement des paramètres de contrôle du modèle.

Nous proposons donc d'adapter au modèle les stratégies articulatoires utilisées par un locuteur en faisant l'hypothèse que ces transitions ont un comportement optimal sur le plan acoustique. Pour cela, nous avons effectué des mesures sur les différents mouvements naturels des articulateurs (lèvres, langue et mâchoire) lors de la réalisation des diphones.

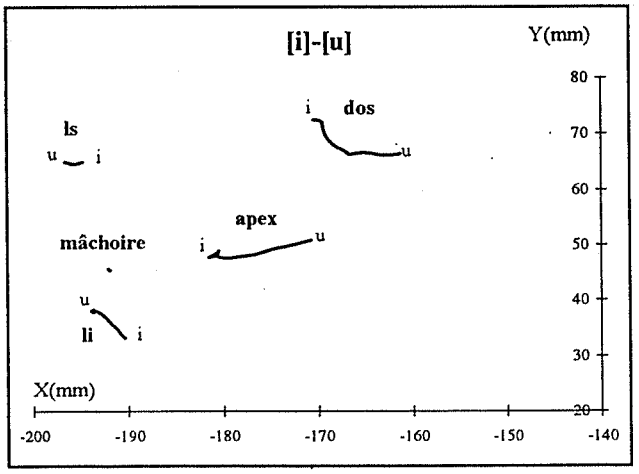


Figure 3: Mouvements des articulateurs (lèvres supérieure et inférieure, incisive inférieure, apex et dos de la langue) du locuteur lc dans le plan X/Y (mm). Les mesures ont été réalisées à l'aide du Movetrack.

3.1. Mesures des mouvements des articulateurs

Les enregistrements ont été effectués pour un locuteur prononçant tous les diphones constitués à partir des 10 voyelles orales du français. Pour cela les capteurs ont été placés sur la lèvre inférieure et sur la lèvre supérieure pour mesurer les mouvements labiaux, sur l'incisive inférieure pour évaluer les variations de la mâchoire et en trois points de la langue (pointe, corps et dos). Dans cette série d'enregistrement, il manque le tracé du corps. Un dernier capteur est positionné sur l'incisive supérieure; il constitue le point de référence permettant la correction éventuelle des données. Ces enregistrements ont été réalisés au LPL (Laboratoire Parole et Langage à Aix-en-Provence) avec un système électromagnétique: le Movetrack (Branderud, 1985; Teston, 1990). Pour chaque séquence V1-V2, nous pouvons visualiser la trajectoire des articulateurs dans le plan X-Y.

3.2. Commande "naturelle" du modèle

Le mouvement des capteurs mesurant l'activité des articulateurs naturels n'est pas directement lié aux paramètres de commande du modèle de Maeda. Les déplacements des capteurs sont projetés sur les axes X et Y. Dans la projection sur l'axe des X les déplacements des articulateurs sont associés aux paramètres du modèle de la manière suivante:

- le mouvement de la lèvre supérieure au paramètre de *protrusion* (lp),
- le déplacement du dernier capteur de la langue au paramètre *position de la langue* (tp).

Dans la projection sur l'axe des Y la correspondance s'effectue comme suit:

- le mouvement de la lèvre inférieure avec le paramètre d'*ouverture labiale* (lh),
- le déplacement de l'incisive avec la commande de la mâchoire (jw),
- le mouvement de la pointe de la langue avec le paramètre de la *pointe de la langue* (tt),
- le déplacement du capteur du centre de la langue avec le paramètre de la *forme de la langue* (ts).

Prenons l'exemple de [i]-[u]: la transition du modèle dont les valeurs des paramètres ont été interpolées linéairement représente très mal la réalité (fig 4). La figure 3 représente le mouvement des articulateurs du locuteur dans le plan X/Y pour la transition [i]-[u]. On constate dans ce cas qu'il n'y a pas de mouvement significatif de la mâchoire pour ce locuteur. Lorsque les déplacements réels des articulateurs ne sont pas suffisamment importants, l'interpolation du paramètre correspondant est effectuée linéairement si nécessaire.

La figure 4 nous permet de comparer l'effet acoustique dans le plan F1/F2 du passage de la configuration du [i] à celle du [u] avec une interpolation linéaire (commande "linéaire") ou déduite des enregistrements des mouvements des articulateurs (commande "naturelle") des paramètres du modèle. La trajectoire caractérisant la production du locuteur constitue la référence idéale dont il convient de se rapprocher au mieux. Nous avons réalisé les mêmes expériences avec les transitions vocaliques [i]-[a], [i]-[y], [a]-[u] et [e]-[u]. Pour toutes ces séquences nous repérons au préalable les articulateurs actifs lors de la transition vocalique naturelle, les autres sont linéairement interpolés entre les deux configurations correspondantes du modèle. Les résultats acoustiques obtenus dans le plan F1/F2 pour les transitions [i]-[u], [i]-[y] et [e]-[u] sont toujours meilleures que les trajectoires formantiques résultant du modèle interpolé linéairement.

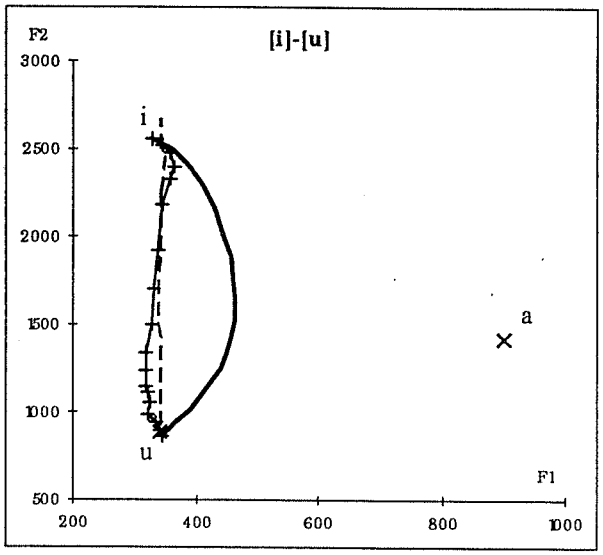


Figure 4: Comparaison de l'effet acoustique dans le plan F1/F2 du passage de la configuration du [i] à celle du [u] avec une commande "linéaire" (trait noir épais) et une commande "naturelle" (caractère "+"). La courbe tracée avec le caractère " - " représente la trajectoire réelle des formants du locuteur lc.

Pour la transition [i]-[a], quelle que soit l'interpolation, toutes les trajectoires du modèles sont équivalentes et représentent très bien la réalité dans le plan F1/F2. Enfin chaque diphone obtenu par cette méthode à été synthétisé et validé perceptivement.

4. CONCLUSION

L'étude du modèle de Maeda a permis de préciser les possibilités qu'offre l'utilisation de ce système de production en vue de la reconnaissance vocale. Une adaptation préalable du modèle aux caractéristiques statiques et dynamiques du locuteur est nécessaire et améliore très significativement les résultats. En effet l'adaptation à plusieurs locuteurs (masculins et féminins) permet d'obtenir des voyelles de référence dont la distance acoustique aux phonèmes du locuteur est améliorée de plus de 70%. La transposition dans le modèle, des vitesses et des accélérations mesurées en divers points des articulateurs, permet de produire des transitions acoustiques très proches de celles du locuteur. Notre stratégie d'adaptation automatique au locuteur, consistant à déterminer une configuration optimale par voyelle, doit être améliorée pour éviter d'obtenir des configurations instables (ie: des formes du conduit vocal dont une faible variation des paramètres articulatoires produit une trop grande variation acoustique). Il convient maintenant d'appliquer ces stratégies contextuelles pour le contrôle dynamique des paramètres du modèle dans un processus de reconnaissance des diphones vocaliques. Des tests sur un plus grand nombre de locuteurs sont envisagés. Enfin cette technique sera appliquée à l'identification de certaines consonnes.

5. BIBLIOGRAPHIE

Boë, L.J., Perrier P. et 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.

Candille L. and Méloni H. (1995) "Automatic speech recognition using production models" ICPhS' 95 Stockholm, vol. 4, 256-259.

Fant G. (1960), *Acoustic theory of speech production*, Mouton, the Hague.

Maddieson, I. (1986), *Patterns of Sounds*. 2nd Edition, Cambridge University Press.

Maeda S. (1979), "Un modèle articulatoire de la langue avec des composantes linéaires", Actes des 10es *JEP*, 154-162.

Payan Y. et Perrier P., (1993), "Vowel normalisation by articulatory normalisation: first attemps for vowel transitions", *Eurospeech 93*, vol.1, 417-420.

Rose R.C., Schroeder J. and Sondhi M.M. (1994), "An investigation of the potential role of speech production models in automatic speech recognition". In Proc. Int. Conf. Sp. Lang. Proc., vol. 2, 575-578.

Teston B., Galindo B. (1990), "Une station de travail d'analyse de la production de la parole", 18es *JEP de la SFA*, Montréal, 28-30 Mai 1990, 180-184.

Vallée N. (1994) *Systèmes vocaliques: de la typologie aux prédictions*, Thèse Doc. ICP Grenoble.

UNE FORMULATION VARIATIONNELLE DU COUPLAGE PHARYNGO-BUCO-NASAL.

Paul JOSPA et Roland VAN PRAAG.
ILVP-Université Libre de Bruxelles (CP110) 50 av. F.D.Roosevelt, 1050 Bruxelles. Belgique.
Tél: 32 2 6503664. Fax: 32 2 6502007. e-mail: pjospa@ulb.ac.be

ABSTRACT:

A variational formulation is given to the acoustic-geometric link for a system of three coupled acoustic tubes. The Rayleigh-Ritz method is used to compute the stationary modes and the transfer function. Compared to classical methods which are based on the transmission line formulation, the variational method permits far faster numerical methods, and facilitates theoretical study.

1. MODES PROPRES.

1.1 Formulation différentielle.

Il est commode d'exprimer le champ sonore par son potentiel des vitesses (Skudrzyk, 1971): $\phi(x,t)$ défini par: $-\partial_x\phi = v(x,t)$, où x est la distance à la glotte, t le temps, et $v(x,t)$ la vitesse volumique du champ sonore; nous exprimons les dérivées indifféremment par les notations $\partial f/\partial x$ ou $\partial_x f$. La pression sonore p peut aussi s'exprimer en terme du potentiel des vitesses: $p(x,t) = \rho\partial_t\phi$, en vertu de l'équation du mouvement: $\partial_x p = -\rho\partial_t v$.

Le système pharyngo-buco-nasal est représenté par trois tubes acoustiques couplés à l'une de leurs extrémités (l'extrémité vélaire). Nous dénoterons par les indices *ph*, *b*, et *n* les grandeurs associées aux conduits respectivement pharyngal, buccal et nasal. Nous adoptons comme origines spatiales: l'extrémité glottique pour le conduit pharyngal, et l'extrémité vélaire pour le conduit buccal et pour le conduit nasal. Pour des raisons de commodité nous utiliserons les variables spatiales suivantes:

$$z_\alpha = x_\alpha / L_\alpha \qquad (1)$$

α désignant l'un des indices *ph*, *b*, ou *n*, et L_α la longueur du tube α. Comme il ne peut y avoir confusion entre les variables z_{ph}, z_b et z_α (elles varient toutes de 0 à 1), nous négligerons d'écrire les indices *ph*, *b*, ou *n* pour désigner celles-ci.

Aux fréquences inférieures à 5000 Hz, on peut monter (Jospa, 1994) qu'à l'intérieur de chacun des tubes, le lien entre les modes propres et la fonction d'aire $A_\alpha(z)$ est exprimé de manière satisfaisante par une équation de Webster:

$$\partial_z(A_\alpha\partial_z\psi_\alpha) + \lambda_\alpha(\omega)A_\alpha\psi_\alpha = 0 \qquad (2)$$

où ψ_α désigne la distribution spatiale d'amplitude du potentiel des vitesses pour le tube α, et avec:

$$\lambda_\alpha(\omega) = (L_\alpha/c)^2\left(\omega^2 - \omega_{p;\alpha}^2\right) \qquad (3)$$

où c désigne la vitesse du son et $\omega_{p;\alpha}$ représente une correction de fréquence pour le tube α traduisant l'effet de l'admittance pariétale sur le champ sonore: $\omega_{p,\alpha} \approx 175\cdot 2\pi s^{-1}$. Un tel lien n'est rigoureusement établi que si l'admittance pariétale est distribuée proportionnellement à la fonction d'aire (Jospa, 1994). C'est précisément le cas du modèle acoustique du conduit vocal proposé par Sondhi (Sondhi, 1974). Les trois potentiels des vitesses sont couplés entre eux en imposant à la jonction des trois conduits, l'égalité des pressions:

$$\psi_b(0) = \psi_n(0) = \psi_{ph}(1) \qquad (4)$$

et la conservation du flux:

$$\frac{A_{ph}(1)}{L_{ph}}\partial_z\psi_{ph}(1) - \frac{A_b(0)}{L_b}\partial_z\psi_b(0) - \frac{A_n(0)}{L_n}\partial_z\psi_n(0) = 0 \qquad (5)$$

Au système différentiel (2)(4)(5), il y a lieu d'adjoindre, dans le cas d'un conduit fermé à la glotte et ouvert aux lèvres et aux narines, les conditions suivantes:

$$\partial_z\psi_{ph}(0) = 0 \quad \text{à la glotte,} \qquad (6)$$

et aux lèvres et aux narines:

$$A_b(L_b)\partial_z\psi_b(L_b) + q_b L_b\sqrt{A_b(L_b)}\psi_b(L_b) = 0 \qquad (7a)$$

$$A_n(L_n)\partial_z\psi_n(L_n) + q_n L_n\sqrt{A_n(L_n)}\psi_n(L_n) = 0 \qquad (7b)$$

avec:

$$q_\alpha = a\sqrt{A(L_\alpha)} + b, \text{ où } a \approx -3.85\text{cm}^{-1}, \text{ et } b \approx 9.8 \qquad (8)$$

Le facteur q_α, déterminé empiriquement, contrôle la partie réactive de l'impédance de rayonnement aux lèvres et aux narines.

1.2 Formulation variationnelle.

Une formulation variationnelle équivalente, qui vaut non seulement pour le conduit ouvert mais également pour le conduit *fermé* aux lèvres, a été donnée à la formulation différentielle du lien acoustico-géométrique exposée ci-dessus (Van Praag & al., 1995). Posons:

$$I_\alpha^0[\psi_\alpha;\omega]=\int_0^1 \frac{A_\alpha}{L_\alpha}\left((\partial_z\psi_\alpha)^2-\lambda_\alpha(\omega)\psi_\alpha{}^2\right)dz \quad (9)$$

avec $\lambda_\alpha(\omega)$ défini par (3) et $\alpha = ph$, b, ou n. On construit trois fonctionnelles en les ψ_α dont les extrémales sont solutions de l'équation de Webster et vérifient les conditions (4-7) imposées à la jonction et aux extrémités du système. Soit:

$$\begin{aligned} &I_{ph}\left[\psi_{ph};\ \omega,\ \psi_b(0),\psi_n(0)\right]\\ &I_b\ \left[\psi_b;\ \omega,\ \psi_{ph}(1),\psi_n(0)\right]\\ &I_n\ \left[\psi_n;\ \omega,\ \psi_{ph}(1),\psi_b(0)\right] \end{aligned} \quad (10)$$

ces fonctionnelles; elles sont définies par:

$$I_{ph}=I_{ph}^0-2\psi_{ph}(1)\left[\frac{A_b(0)}{L_b}\partial_z\psi_b(0)+\frac{A_n(0)}{L_n}\partial_z\psi_n(0)\right] \quad (11a)$$

$$I_b=I_b^0+2\psi_b(0)\left[\frac{A_{ph}(1)}{L_{ph}}\partial_z\psi_{ph}(1)-\frac{A_n(0)}{L_n}\partial_z\psi_n(0)\right]+ \psi_b(0)\left[2\psi_{ph}(1)-\psi_b(0)\right]L_b+q\sqrt{A_b(1)}\psi_b^2(1) \quad (11b)$$

$I_n = I_b$ en inversant les indices b et n. (11c)

Les extrémales ψ_α de ces fonctionnelles s'identifient aux distributions d'amplitude des modes. Elles sont obtenues en résolvant le système variationnel suivant (Courant & al., 1953):

$$\delta I_\alpha/\delta\psi_\alpha = 0,\ (\alpha = ph,b,n) \quad (12)$$

Ce système étant homogène, il n'admet de solution que pour des valeurs discrètes de la fréquence ω (fréquences des modes propres). La méthode de Rayleigh-Ritz (Courant & al., 1953) fournit un moyen efficace pour résoudre directement le système variationnel à l'approximation désirée (voir annexe A.1).

2. FONCTION DE TRANSFERT.

Nous exprimons les potentiels des vitesses $\phi_\alpha(z,t)$ à l'aide de leurs transformées de Fourier:

$$\phi_\alpha(z,t)=\int_{-\infty}^{\infty}\psi_\alpha(z;\omega)e^{j\omega t}d\omega\ .$$

Sous les mêmes hypothèses que précédemment, nous pouvons admettre que pour toute fréquence $\omega<5000$hz l'amplitude $\psi_\alpha(z;\omega)$ vérifie l'équation de Webster (2). Les conditions de jonction (4) et (5) ainsi que les conditions (7) aux lèvres et aux narines restent également inchangées.

La fonction de transfert $H(x_\alpha;\omega)$ en un point quelconque x_α du système est définie par le rapport $\left|v_\alpha(x_\alpha;\omega)/v_{ph}(0;\omega)\right|$. En adoptant, en lieu et place de (6), la condition suivante à la glotte:

$$\partial_z\psi_{ph}(0;\omega)=1, \quad (13)$$

la fonction de transfert sera donnée directement par:

$$H(x_\alpha;\omega)=\frac{L_{ph}}{L_\alpha}\left|\partial_z\psi_\alpha(z;\omega)\right|. \quad (14)$$

en raison de la définition du potentiel des vitesses.

Le calcul de la fonction de transfert peut également recevoir une formulation variationnelle. Pour ce faire, nous introduisons ici encore trois fonctionnelles en les ψ_α dont les extrémales sont solutions de l'équation de Webster (2) et vérifient les conditions (4),(5),(7) et (13). Comparées aux fonctionnelles (11), seule diffère la fonctionnelle I_{ph} relative au conduit pharyn-gal: un terme défini à la glotte lui est ajouté de manière à ce que son extrémale vérifie nécessai-rement la condition (14); ceci nous donne:

$$I_{ph}=I_{ph}^0-2\psi_{ph}(1)\left[\frac{A_b(0)}{L_b}\partial_z\psi_b(0)+\frac{A_n(0)}{L_n}\partial_z\psi_n(0)\right]+ 2\left(A_{ph}(0)/L_{ph}\right)\psi_{ph}(0). \quad (15)$$

Ici encore, la solution au problème s'obtient en résolvant le système (12). De par la présence du terme linéaire en la fonction ψ_{ph} dans I_{ph}, le système (12) n'est plus homogène en les fonctions ψ_α. Ce système présente donc maintenant une solution quelque soit la valeur de le fréquence ω.

La méthode de Ritz permet, ici encore, d'appro-cher efficacement les potentiels des vitesses désirés en tant qu'extrémales des fonctionnelles (11a,b) et (15) couplées par les conditions de jonction (voir annexe A.2). Finalement, la fonction de transfert est obtenue par (14).

3. COMPARAISONS ET EXEMPLE.

3.1 Temps de calcul.

Le calcul des modes par la méthode variationnelle permet une économie de temps considérable. Pour évaluer ce gain, nous avons comparé notre méthode (méthode VAR) avec la méthode classique de Liljencrants et Fant (Liljencrants & al., 1975) (méthode LF), pour le calcul des fréquences des 4 premiers modes dans *un tube unique* (figure 1). La méthode LF est basée sur le calcul des matrices de transfert d'une ligne de transmission sans pertes.

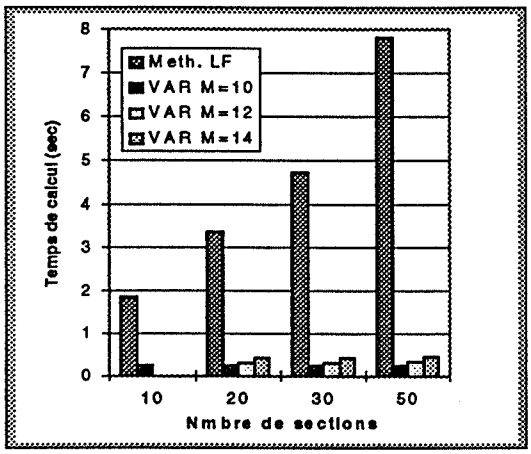


Figure 1: Temps de calcul des 4 premières fréquences. propres d'un tube unique sans pertes, en fonction du nombre de sections du tube. VAR M=10 désigne la méthode variationnelle avec 10 fonctions coordonnées. Programmes MATLAB implantés sur 486DX2 80Mhz.

3.2 Mesures de J.Dang et K. Honda.

Sur un système de trois tubes cylindriques couplés, Dang et Honda (Dang & al., 1994) ont effectués des mesures précises de la fonction de transfert entre un point intérieur x_0 au système (dans le tube *ph* selon notre notation) et l'extrémité ouverte (rayonnante) du tube *b*, soit $\left|v_b(L_b;\omega)/v_{ph}(x_0;\omega)\right|$. Le tube *n* est fermé à son extrémité libre. La fonction de transfert calculée par notre méthode (la position des zéros en particulier: 1375, 4125, 6883 hertz) est en excellent accord avec celle mesurée par Dang et Honda.

3.3 Exemple.

Pour illustrer la méthode, considérons la configuration suivante: $L_{ph}=8.5cm$, $D_{ph}=2.2cm$, $L_b=8cm$, $D_b=2.2cm$, $L_n=10cm$, $D_n=1.6cm$.; les tubes *b* et *n* sont ouverts à leurs extrémités libres, le tube *ph* est fermé à l'entrée. Les fréquences calculées des modes sont: 596, 1440, 2158, 3355, 4294, 4935, 6128, 7078 hertz. Les figures 2-3 représentent les distributions d'amplitude $\psi_\alpha(x_\alpha)$ pour les modes aux fréquences 1440hz et 2158hz. On constate un comportement qualitatif différent entre ces deux modes. Pour le mode 2, les distributions d'amplitude dans les branches buccale et nasale sont en phase, et les admittances d'entrée de ces deux branches sont de même signe. On observe un comportement inverse pour le mode 3. On s'attend à ce que ce dernier comportement caractérise les formants propres à la nasalité qui interviennent dans les paires pôle-zéros. Ceci se trouve confirmé par le calcul de la fonction de transfert (voir figure 4).

ANNEXE

A.1 Calcul des modes propres.

On choisi un ensemble $\{\eta_i(z)\}$ de M fonctions linéairement indépendantes (fonctions coordon-nées) définies sur $[0,1]$. Les solutions ψ_α du problème sont approchées par des combinaisons linéaires de la forme:

$$\begin{aligned}\psi_{ph} &= \sum_{i=1}^{M} C_i \eta_i(z),\\ \psi_b &= \sum_{i=1}^{M} C_{i+M} \eta_i(z),\\ \psi_n &= \sum_{i=1}^{M} C_{i+2M} \eta_i(z)\end{aligned} \qquad (16)$$

où, pour simplifier l'écriture, nous avons choisi le même nombre M de fonctions coordonnées par conduit. Après substitution de (16) dans les fonc-tionnelles (11), et après intégration sur la variable spatiale *z*, ces fonctionnelles deviennent des fonc-tions quadratiques ordinaires en les coefficients C_i $(i=1,...,3M)$. La recherche des extrémales des fonctionnelles (11) est ainsi ramenée à un problème classique de minimisation de $3M$ fonc-tions quadratiques en les $3M$ variables C_i. Nous obtenons ainsi un système de $3M$ équations liné-aires et homogènes à $3M$ inconnues de la forme:

$$\sum_{j=1}^{3M}\left(K_{ij}-\omega^2 V_{ij}\right)C_j = 0 \quad (i=1,...,3M) \qquad (17)$$

où les éléments des matrices K et V sont des nombres obtenus par intégration sur z dans les expressions des fonctionnelles (11), (les ψ_α étant approchés explicitement par (16)) Pour un choix donné de fonctions coordonnées, ces éléments ne dépendent que des paramètres physiques des conduits (fonctions d'aire, corrections de fréquence liée à l'admittance des parois,

etc.). Ce système n'a de solutions non triviales que si ω^2 est valeur propre de la matrice $V^{-1}K$. Ainsi, une fois construites les matrices K et V, la recherche des fréquences ω (fréquences formantiques) et des distributions spatiales ψ_α des modes, se réduit à l'application d'une méthode classique de calcul des valeurs et vecteurs propres d'une matrice.

Un programme effectuant le calcul des modes selon cette méthode à été réalisé. Comme fonctions coordonnées η_i nous avons choisi les M premiers polynômes de Tchebychev $T_i(2z-1)$ en raison de leur comportement aux limites. Une précision satisfaisante est obtenue pour M=7.

A.2 Calcul des fonctions de transfert.

Le même procédé que celui présenté en A.1 est appliqué à la recherche des extrémales des fonctionnelles (15) et (11b,c). En lieu et place de (17), nous obtenons à présent un système linéaire *non* homogène en les coefficients C_j de la forme:

$$\sum_{j=1}^{3M}\left(K_{ij}-\omega^2 V_{ij}\right)C_j = R_j \qquad (i=1,\ldots,3M) \qquad (18)$$

où, pour un choix donné de fonctions coordonnées, les éléments du vecteur R ne dépendent que du rapport: $A_{ph}(0)/L_{ph}$. Les C_j, fonctions de ω, sont donc donnés par: $C(\omega)=\left(K-\omega^2 V\right)^{-1}R$. Il suffit dès lors d'introduire les $C_j(\omega)$ ainsi calculés dans les ψ_α pour que les fonctions de transfert données par (14) soient immédiatement calculables.

Références:

Skudrzyk E. (1971) *The Fundation of Acoustics.* Springer-Verlag, Wien, New-York.

Jospa P. (1994) Formulation variationnelle du lien acoustico-articulatoire. *Actes des 20èmes J.E.P.*, Trégastel, France, pp.113-118

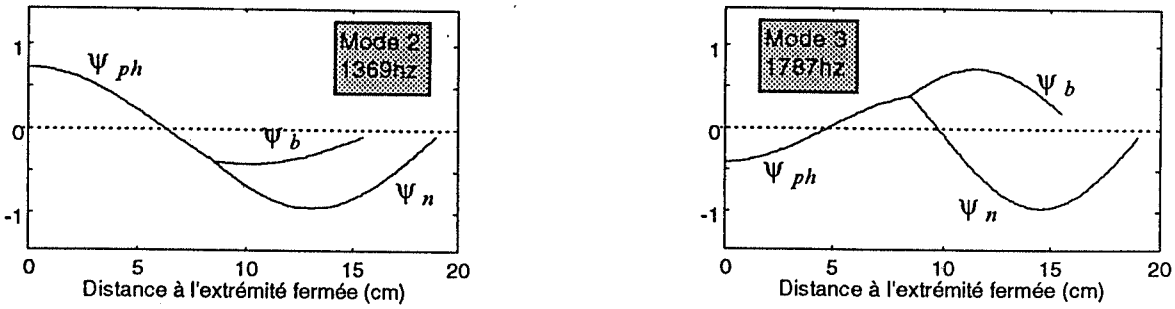
Sondhi M. M. (1974) Model for wave propagation in a lossy vocal tract. *J. A. S. A.*, Vol.55, n°5, 1070-1075.

Van Praag R. & Jospa P (1995) Variational method applied to formant computation for a pharyngo-buco-nasal tract. *Proc. ICPhS 95*, Stockholm, V 4,.456-459.

Courant R. & Hilbert D. (1953) *Methods of mathematical physics.* Interscience Publ. Inc., Vol. 1.

Dang J. & Honda K. (1994) A new method for measuring vocal tract transmission characteristics. *Tech. Report of ATR*, TR-H-108.

Liljencrants J. & Fant G (1975) Computer program for VT-resonance frequency calculations. STL-QPSR 4/1975, pp15-21.



Figures 2-3: Potentiels des vitesses (ou amplitudes des pressions) pour deux modes stationnaires dans les trois branches du système décrit au.§3.2. Le mode 3 correspond à un extra-formant appartenant à une paire pôle-zéro induite par le couplage de la branche *n* .

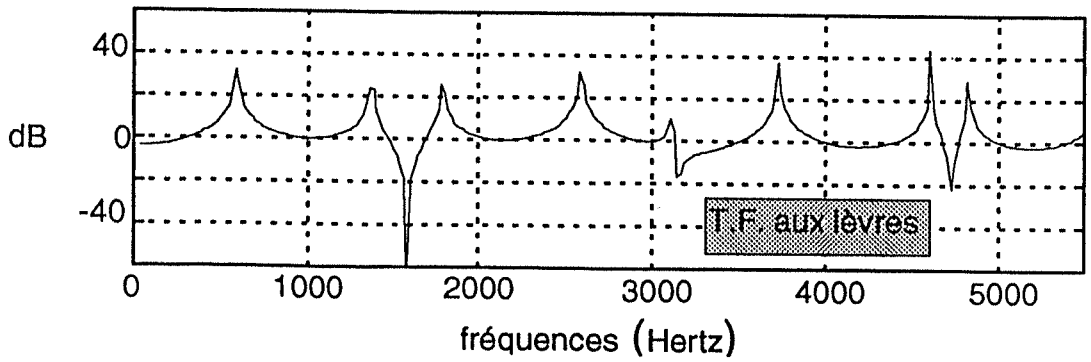
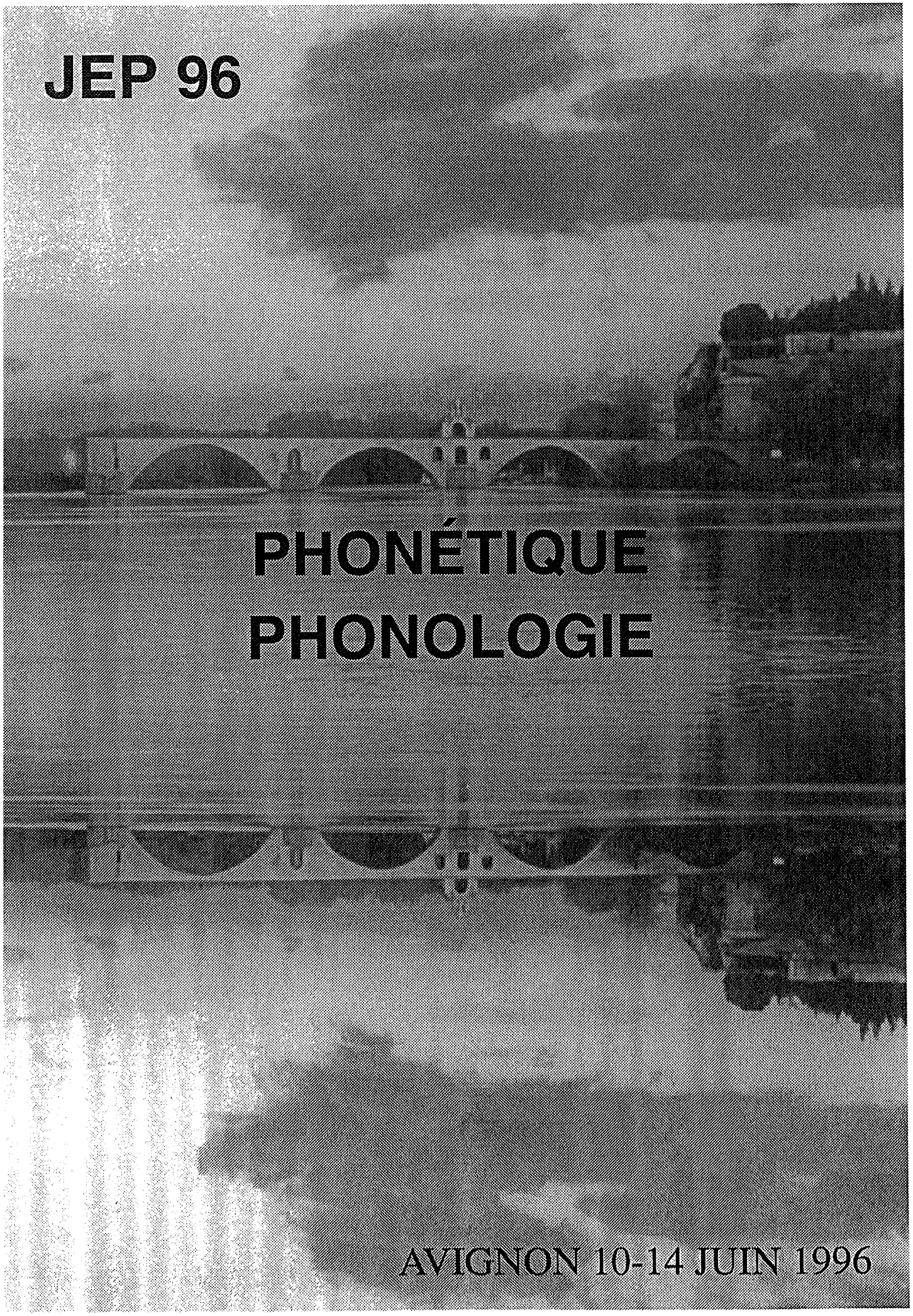


Figure 4: Fonction de transfert de l'origine du tube *ph* ("glotte") à l'extrémité ouverte du tube *b* ("lèvres"). Système acoustique de même configuration que pour les figures 2-3 (voir texte §3.2).

JEP 96



PHONÉTIQUE
PHONOLOGIE

AVIGNON 10-14 JUIN 1996

ASPECTS AERODYNAMIQUES ET ARTICULATOIRES DES OCLLUSIVES LABIO-VELAIRES

Didier DEMOLIN et Bernard TESTON

Université libre de Bruxelles et Université d'Aix-en-Provence-URA 261 CNRS
Université Libre de Bruxelles-CP 175, 50 av. F. Roosevelt 1050 Bruxelles
Tél.: ++ 32 2 650 45 07 - Fax: ++ 32 2 650 24 50 -email: ddemoli@ulb.ac.be

ABSTRACT

This paper examines labio-velar stops found in three African languages, Mangbetu, Liko and Lendu. Focus is on the aerodynamic parameters of pharyngeal pressure and oral flow. Articulatory data consist of an examination of video images showing the lips and lower jaw movements. Pharyngeal pressure plots suggest that there is an important articulatory variabilty in the realization of labio-velars, involving front and back movements of the tongue body and sometimes a lowering of the larynx. Video images clearly show that the lower jaw movement involved in the production of labio-velars is different than the movement made to produce single velar or bilabial stops even if they are implosive.

1. INTRODUCTION

Les occlusives labio-vélaires sont parmi les sons à double articulation les plus répandus dans les langues du monde. On les rencontre dans plusieurs familles linguistiques différentes et en particulier dans les langues d'Afrique centrale. Les consonnes labio-vélaires se rencontrent dans deux des principales familles linguistiques de l'Afrique, le Niger-Kordofan et le Nil-Sahara. L'étude la plus importante sur ce type de consonnes a été faite par Connell (1994) qui les a examinées dans cinq langues du Nigéria en utilisant différentes techniques instrumentales (spectrographie, aérométrie, laryngographie et électropalatographie).

Dans cet article, nous examinerons les labio-vélaires du liko, une langue bantoue appartenant à la famille Niger-Kordofan ainsi que celles du mangbetu et du lendu, deux langues du groupe Soudan central appartenant à la famille Nil-Sahara. Pour chaque langue, les données ont été acquises avec un locuteur. L'accent est mis sur les paramètres aérodynamiques de pression pharyngale et de débit oral. Les données articulatoires consistent en un examen du mouvement des lèvres et de la mâchoire inférieure, au moyen d'images vidéo prises simultanément de face et de profil. L'examen de la pression pharyngale, dans les langues examinées, suggère qu'il existe une variabilité articulatoire importante dans la réalisation des labio-vélaires, que ce soit entre les langues ou entre la réalsation des sourdes et des sonores. L'examen des images vidéo compare la réalisation des labio-vélaires avec celles des occlusives bilabiales et des vélaires simples, dans le but de décrire le mouvement de la mâchoire inférieure et de formuler une hypothèse sur le mouvement que réalise la langue pendant l'articulation de ces consonnes.

Le statut phonologique des labio-vélaires du mangbetu est donné dans Demolin (1991, 1992); celui du lendu notamment dans Dimmendaal (1986) et Mertens (1987); celui du liko dans Demolin (à paraître).

2. DONNEES AERODYNAMIQUES

L'examen du débit oral en mangbetu montre qu'au moment du relâchement de l'occlusion, tant avec les sourdes qu'avec les sonores, il existe un débit négatif. Ce débit oral négatif se retrouve également en lendu et en liko (figure 1) ainsi que dans les mesures de Ladefoged (1964) et de Connell (1994), il est plus important après les sourdes qu'après les sonores et reflète un mécanisme d'initiation vélaire. Ce mécanisme peut être décrit comme suit: après la réalisation des deux occlusions, il y a un mouvement descendant de la mâchoire, et un mouvement vers l'arrière du point de contact de la langue avec le palais mou. Ces mouvements ont pour effet de provoquer une chute de pression à l'intérieur de la bouche. A ce moment, la fermeture labiale se relâche avant la vélaire, de sorte qu'une occlusive ingressive, produite par succion vélaire, est réalisée. Les mouvements précis des articulateurs sont hypothétiques, cependant la similarité des données aérodynamiques et acoustiques avec d'autres descriptions basées sur des données cinéradiographiques nous permet de penser que les mêmes mouvements sont à l'oeuvre ici.

La pression pharyngale présente, quand à elle, des comportements différents lorsqu'on compare le mangbetu au liko et au lendu (figures 1, 2 et 3). En mangbetu, avec la sourde [kp], on observe une brusque montée de la pression, immédiatement après la fermeture

vélaire, ensuite une montée très progressive pendant le reste de l'occlusion suivie d'une chute très rapide, au moment du relâchement. Le phénomène est identique pour la sonore [g b], avec cependant une augmentation de pression nettement moins importante, conséquence du débit oral plus faible lors de la réalisation des occlusives sonores. En liko, la sourde [k p] montre un comportement différent. Après la brusque montée de pression qui suit la première occlusion, la pression se maintient mais avec une diminution très progressive, jusqu'à la chute rapide de la pression au moment du relâchement. De plus, on peut observer une pression pharyngale négative pendant un moment, après le relâchement. Ce phénomène peut s'observer dans tous les cas avec les sourdes, il est variable et peu marqué avec les sonores. La diminution progressive de la pression pendant l'occlusion semble indiquér un mouvement vers l'avant de la langue pendant l'occlusion. La pression négative après le relâchement semble quant à elle indiquer, soit une fermeture de la glotte pendant l'occlusion, soit une descente du larynx comme cela a été suggéré par Ladefoged (1964) et par Connell (1994). En lendu, le comportement des labio-vélaires est similaire à celui qu'on observe en liko. La pression pharyngale négative après le relâchement est cependant moins longue qu'en liko avec les sourdes, elle est peu marquée ou inexistante avec les sonores. Le tableaux 1, 2 et 3 résument les données de débit oral et de pression pharyngale immédiatement après le relâchement de la dernière occlusion, pour les trois langues étudiées.

Tableau 1.Résumé des paramètres de débit oral et de pression pharyngale.

	mangbetu	*lendu*	*liko*
Débit oral	< 0	< 0	< 0
Pression pharyngale	≥ 0 après [kp] ≥ 0 après [gb]	< 0 après [kp] ≥ 0 ou < 0 après [gb]	< 0 après [kp] < 0 après [gb]

Tableau 2. Débit oral moyen en l/m au relâchement des labio-vélaires.

	mangbetu	*lendu*	*liko*
sourde [kp]	- 13 l/m	- 9 l/m	- 10 l/m
sonore [gb]	- 10 l/m	- 6 l/m	- 8 l/m

Tableau 3. Pression pharyngale moyen en Hpa au relâchement des labio-vélaires.

	mangbetu	*lendu*	*liko*
sourde [kp]	0	- 4.5 Hpa	- 10 Hpa
sonore [gb]	0	0	- 0.1 Hpa

3. DONNEES ARTICULATOIRES

Les données articulatoires prises en vidéo ont été obtenues avec un locuteur lendu. L'examen a consisté à acquérir, au moyen d'une caméra vidéo, des images de face et de profil de l'articulation des consonnes labio-vélaires de cette langue. L'ouverture au 1/1000 de la caméra permet d'obtenir une image toutes les 20 ms. Le protocole consistait à observer le mouvement de la mâchoire inférieure pour les labio-vélaires [k p] et [g b], comparées avec les occlusives vélaires [k] et [g], avec les occlusives bilabiales [p] et [b] ainsi qu'avec l'implosive [ɓ].

Ces images permettent d'observer que pendant l'occlusion des labio-vélaires, il y a un mouvement descendant de la mâchoire inférieure qui est nettement perceptible. Ce mouvement n'est pas visible avec les occlusives simples, qu'elles soient vélaires ou bilabiales. Le mouvement qui est visible au relâchement des occlusives [k, g, p, b] est descendant mais avec une trajectoire d'avant en arrière alors que celui des labio-vélaires va de haut en bas. Le seul mouvement comparable est celui de l'implosive [ɓ], mais dans ce cas, en plus du mouvement descendant de la mâchoire pendant l'occlusion, il y a un mouvement de retraction de la langue qui n'est pas perceptible pour les labio-vélaires.

La figure 4 montre la position de la mâchoire inférieure au début et à la fin de l'occlusion dans la première série de mots.

4. DISCUSSION

Le schéma donné à la figure 5 illustre le mouvement suggéré pour décrire la succion vélaire, telle que nous l'avons décrite en 2 et telle qu'elle a été proposée par Ladefoged (1964), Demolin (1992) et Connell (1994) pour expliquer l'initiation des labio-vélaires. Cependant, Catford (1977) a soumis une autre explication possible pour rendre compte du débit oral négatif, qui est observé au relâchement de l'occlusion des labio-vélaires. Selon lui, la brusque augmentation du volume de la cavité buccale, après le relâchement de l'occlusion pourrait être à l'origine de cette chute du débit oral. Il existe une troisième explication, si l'on peut montrer que le larynx descend pendant l'occlusion. Le débit oral négatif est alors une conséquence de la pression pharyngale négative qui est générée par la descente du larynx, ce qui semble être le cas en liko et en lendu, au moins pour les sourdes. Le mouvement de la langue suggéré à la figure 5 semble être en contradiction avec les données de pression pharyngale observées avec [kp] en liko et en lendu, puisque nous suggérons une légère avancée de la langue pendant l'occlusion. Dans ce cas, la chute de pression à l'intérieur de la bouche ne peut plus être une conséquence du recul de la masse de la langue à l'intérieur de la cavité buccale. La variation observée dans les tracés de pression pharyngale, entre les différentes langues étudiées, et la similarité des tracés de débit oral doit en outre être expliquée, si l'on veut donner une description cohérente de l'articulation de ces consonnes. L'explication est de faire l'hypothèse d'un abaissement progressif de la langue pendant l'occlusion, accompagné soit d'un recul, soit d'une légère avancée de la masse de la langue. Ce mouvement est suivi d'un abaissement rapide au moment du relâchement. Ce mouvement de la langue peut en outre être accompagné d'un abaissement du larynx qui rend compte de la pression pharyngale négative observée au relâchement de l'occlusion de [kp] en liko et en lendu. Cette pression négative indique que le larynx n'a pas encore repris sa position ou moment du relâchement. La pression normale ne se rétablissant en moyenne que 15 ms après le relâchement en liko et 10 ms en lendu.

5. CONCLUSION

L'examen des données aérodynamiques des occlusives labio-vélaires du mangbetu, du lendu et du liko montre qu'il existe une certaine variabilité dans la réalisation de ces consonnes. Cette variabilté peut être décrite et comprise en examinant les paramètres de pression pharyngale et de débit oral. Le mouvement d'abaissement de la mâchoire inférieure, différent de celui d'occlusives vélaires et bilabiales simples est un mouvement d'abaissement, qui est perceptible dès le début de la première occlusion.

5. BIBLIOGRAPHIE

Catford I.. 1977. *Fundamental Problems in Phonetics.* Edinburgh University Press.

Connell B. 1994. The stucture of labial-velar stops. *Journal of Phonetics 22*, 441-476.

Demolin D. 1991. Les consonnes labio-vélaires du mangbetu. *Pholia 6*, 85-105.

Demolin D. 1992. *Le mangbetu: étude phonétique et phonologique.* Thèse de doctorat. Université libre de Bruxelles.

Demolin D. Phonétique et phonologie du liko. A paraître dans Pholia 10.

Dimmendaal G. J. 1986. Language Typology, Comparative linguistics, and injective Consonants in Lendu. *Afrika und Übersee, Band 69*, 161-192.

Ladefoged P. 1964. *A Phonetic Study of West-African languages: an auditory and instrumental survey.* Cambridge, C.U.P.

Ladefoged P. et I. Maddieson. 1995. *The Sounds of the World's Languages.* Oxford. Blackwell.

Maddieson I. et P. Ladefoged. 1989. Multiply articulated segments and the feature hierarchy. *UCLA Working Papers in Phonetics 72*, 116-138.

Mertens. 1987. *Dictionnaire bba-dha.* Ddrøddrø.

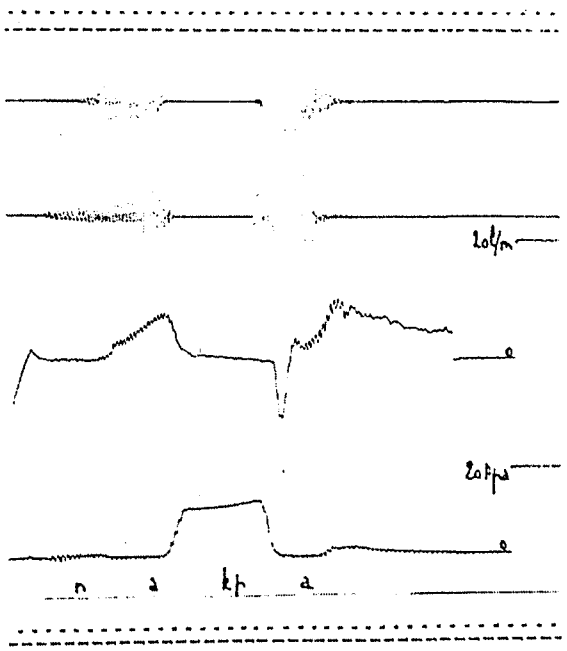


Figure 1. Signal (1), eléctroglottogramme (2), débit oral (3) et pression pharyngale (4) pour le mot n à k p á 'esclave' en mangbetu.

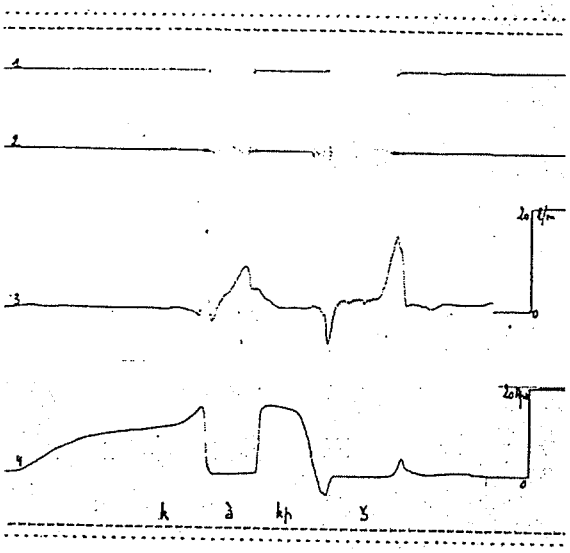


Figure 2. Signal (1), eléctroglottogramme (2), débit oral (3) et pression pharyngale (4) pour le mot k à k p ɔ 'creuser' en liko.

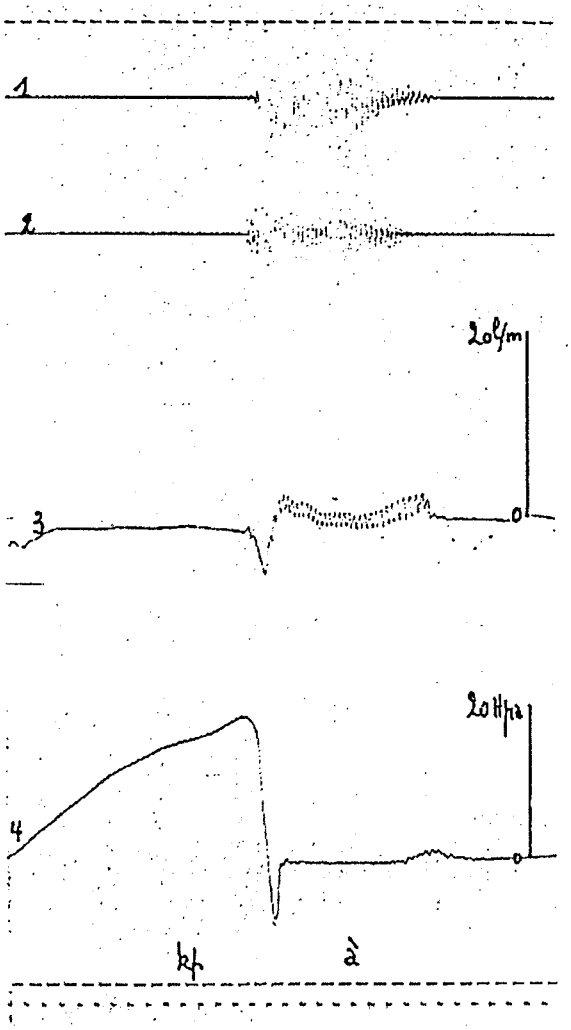


Figure 3. Signal (1), eléctroglottogramme (2), débit oral (3) et pression pharyngale (4) pour le mot k p à 'couper' en lendu

(a)

(b)

Figure 4. Position des lèvres et de la mâchoire inférieure au début de l'occlusion et juste après le relâchement de [kp] (a et b).

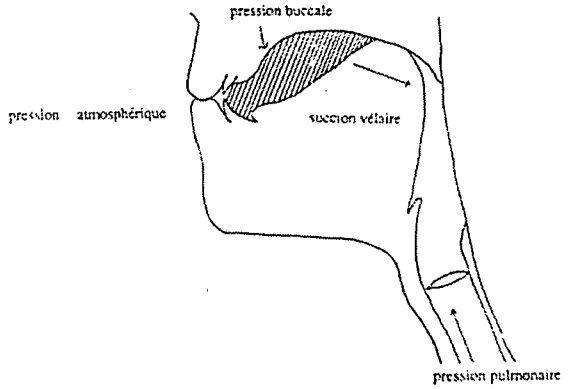


Figure 5. Schéma du mouvement articulatoire à l'origine de la succion vélaire.

CARACTERISTIQUES PHONETIQUES DU SYSTEME VOCALIQUE DU BOBO-FING

Annelies BULKENS et Didier DEMOLIN

Université Libre de Bruxelles-CP 175, 50 av. F. Roosevelt 1050 Bruxelles
Tél.: ++ 32 2 650 45 07 - Fax: ++ 32 2 650 24 50 -email: ddemoli@ulb.ac.be

ABSTRACT

This paper describes the main characteristics of the vowels found in Bobo-Fing, a mande language spoken in Burkina-Faso. Bobo-Fing has an interessting vowel system including 10 oral vowels and 5 nasal vowels. An acoustic study is made to compare formant frequencies of the oral vowels and their nasal conterpart. Results are discussed to show the main differences between both sets of vowels. Finally the influence of nasal vowels on neighbouring segments is examined. Epenthetic processes involving the small velar closure found at the end of back nasal vowels are discussed.

1. INTRODUCTION

Le bobo-fing (ou madarɛ) est une langue mande parlée au Burkina-Faso, principalement dans la région autour de la ville de Bobo-Dioulasso. Elle est parlée par environ 300 000 locuteurs. La classification des langues mande a longtemps considéré le bobo-fing comme une branche autonome (Welmers 1958; Long 1971), mais des données plus récentes rapprochent plutôt cette langue de la subdivision mande-nord (Grégoire-de Halleux 1994). Les caractéristiques des langues mande sont, entre autres, une structure syntaxique SOV, l'absence de classes nominales, un système tonal et une morphologie assez simple ainsi qu'une tendance marquée à utilIser des mots monsyllabiques. Le bobo-fing, comme beaucoup de langues du groupe mande (voir par exemple Ruhlen 1978 et Maddieson 1984) possède un ensemble de voyelles nasales. Cet article, montre le statut phonologique des voyelles orales et nasales du bobo-fing et compare ensuite les caractéristiques acoustiques de ces voyelles. Les données ont été acquises avec un locuteur masculin.

2. LE SYSTEME VOCALIQUE DU BOBO-FING

Le bobo-fing a deux séries de voyelles. Une série de voyelles orales qui compte dix phonèmes /i, ɪ, e, ɛ, a, ə, ɔ, o, ʊ, u/ et une série de nasales qui compte cinq phonèmes /ĩ, ɛ̃, ã, ɔ̃, ũ/. Le statut phonologique de ces voyelles resort des quelques exemples qui suivent.

d à	langue	d ò	l'âge
d ì	en premier lieu		
b é	là	b í	tu, toi
b ó	ce		
k á	mettre	k ó	rentrer
k ò	cinq	k ù	marigot
s ʊ̀mà	aujourd'hui	s ə̀má	les gens
l ú	maison	l ó	sol
l é	toucher		
kwɛ̃	maisons	kwã	peur
kɔ̃	marché	kũ	sorte d'arbre
sɔ̃	homme	s ɔ̀	couper
sɛ̃	serpent	sĩ	l'homme
sĩ	soleil	s í	race

Notons que les voyelles nasales [ɪ̃] et [õ] sont aussi réalisées, mais seulement comme variantes phonétiques de [ĩ] et de [ɔ̃].

3 COMPARAISON ENTRE VOYELLES ORALES ET NASALES

Les principales différences entre voyelles orales et voyelles nasales sont: la modification spectrale due au couplage des cavités nasales qui peut être interprétée comme l'adjonction de paires formant-antiformant, F_{in} et A_{in}, leur position fréquentielle dépendant du degré d'abaissement du voile du palais. Le couplage entre la cavité orale et la cavité nasale provoque en outre une modification des fréquences liées au seul conduit oral, ces formants sont notés F_i'.

Chaque paire de voyelles, l'orale et la nasale correspondante [i]-[ĩ], [ɛ]-[ɛ̃], [a]-[ã], [ɔ]-[ɔ̃], [u]-[ũ] sont comparées du point

de vue acoustique. Un corpus incluant l'ensemble des voyelles orales et nasales a été enregistré. Les formants des voyelles ont été mesurés au moyen du logiciel Phonédit. Pour chaque segment les mesures ont été faites à l'instant qui concilie le plus fort niveau sonore avec la plus forte stabilité spectrale. Chaque voyelle a été analysée avec le spectre prélévé a cet instant du segment.

La paire [i]-[ĩ]

Les principales différences entre l'orale et la nasale sont une montée d'environ 100Hz de F_1', F_2' monte également d'environ 210 Hz. Il n'y a pas le mouvement de compacité croissante entre F_1' et F_2' auquel on peut s'attendre. F_3' est toujours plus haut d'environ 390 Hz.

La paire [ɛ]-[ɛ̃]

Les principales différences entre l'orale et la nasale sont une montée d'environ 120Hz de F_1', F_2' descend d'environ 140 Hz, soit pour les deux formants un mouvement de compacité croissante entre F_1' et F_2'. F_3' est toujours plus haut d'environ 170 Hz.

La paire [a]-[ɑ̃]

Les principales différences entre l'orale et la nasale sont une descente d'environ 140Hz de F_1', F_2' descend d'environ 390 Hz, il n'a donc pas de mouvement de rapprochement des deux formants pour la nasale. F_3' est aussi toujours plus bas d'environ 960 Hz. Le spectre de [ɑ̃] est plus atténué que celui de [a].

La paire [ɔ]-[ɔ̃]

Les principales différences entre l'orale et la nasale sont une montée d'environ 45Hz de F_1', F_2' descend d'environ 110 Hz, soit pour les deux formants un mouvement de compacité croissante entre F_1' et F_2'. F_3' est toujours plus bas d'environ 520 Hz.

La paire [u]-[ũ]

Les principales différences entre l'orale et la nasale sont une montée d'environ 40Hz de F_1', F_2' descend d'environ 640 Hz, soit pour les deux formants un mouvement de compacité croissante entre F_1' et F_2'. F_3' est toujours plus haut d'environ 225 Hz.

A l'exception de ce qui s'observe pour la paire [a]-[ɑ̃], F_1' est toujours plus haut pour la voyelle nasale, mais de manière moins marquée pour les voyelles postérieures que pour les voyelles antérieures. L'hypothèse selon laquelle F2 ne serait pas affecté par la nasalisation (Calliope 1989) n'est pas justifiée par les données que l'on peut observer en bobo-fing. A l'exception de ce qu'on constate avec la paire [i]-[ĩ], F_2' est toujours plus bas pour les voyelles nasales, cette différence étant particulièrement marquée pour la voyelle postérieure fermée [u]. L'augmentation de fréquence de F_3' qui est décrite comme un bon indice de nasalité (Calliope 1989) n'est pas toujours observée. Elle n'existe pas dans nos données pour les paires [a]-[ɑ̃] et [ɔ]-[ɔ̃], dans lesquelles le F_3' de la voyelle nasale est nettement plus bas que celui de la voyelle orale.

On peut constater d'une manière générale, la présence d'un premier formant formant nasal Fn1 autour de 260 Hz. La présence d'un second formant nasal Fn2 vers 2000 Hz est aussi constatée sauf pour les voyelles [ɑ̃] et [ɔ̃].

Notons encore que la voyelle orale la plus proche de [ɑ̃] est [ɔ]. La voyelle orale la plus proche de [ĩ] est [ɪ], et la voyelle la orale la plus proche de [ɔ̃] est [ʊ]. Les tableaux 1 et 2 donnent les valeurs moyennes pour F1, F2 et F3 des voyelles orales [i, ɛ, a, ɔ, u] et F_1', F_2', F_3', pour les voyelles nasales correspondantes.

Table 1. Fréquence de F1, F2, F3 en Hz pour les voyelles orales [i, ɛ, a, ɔ, u].

	F1	F2	F3
[i]	258	2308	3031
[ɛ]	430	2049	3435
[a]	769	1748	3693
[ɔ]	576	1213	3203
[u]	324	1638	3515

Table 2. Fréquence en Hz de F_1', F_2', F_3' pour les voyelles nasales [ĩ, ɛ̃, ɑ̃, ɔ̃, ũ].

	F1'	F2'	F3'
[ĩ]	358	2530	3422
[ɛ̃]	550	1909	3614
[ã]	629	1353	2726
[ɔ̃]	622	1200	2679
[ũ]	364	1034	3740

4. CENTRE DE GRAVITE

Chistovitch et Lublinskaya (1979) et Chistovitch (1979) on montré que la perception du degré d'aperture des voyelles reflète un 'centre de gravité' déterminé par la fréquence et l'amplitude des pics spectraux dans le région de F1-F2. Nous avons mesuré le centre de gravité des voyelles orales et nasales en calculant la fréquence moyenne de la zone comprise sous l'enveloppe spectrale de la région F1-F2, selon une méthode proposée par Beddor (1982). Ces mesures montrent que le centre de gravité est plus haut pour [ĩ]-[ɛ̃] que pour [i]-[ɛ] et plus bas pour [ã]-[ɔ̃] que pour [a]-[ɔ]. Les mesures que nous avons faites pour [ũ] montrent qu'il le centre de gravité de cette voyelle est légèrement plus haut que celui de [u].

5. LES VOYELLES POSTERIEURES

Les voyelles nasales du bobo-fing, lorsqu'elles sont en fin de mot, ont souvent une brève fermeture vélaire [ŋ]. Cette fermeture vélaire est similaire à celle que l'on peut entendre dans le français méridional, à la fin de mots tels que [bɑ̃ŋ]. Des exemples presque identiques ont été cités par Ruhlen (1978). Ohala et Ohala (1993) citent d'autres exemples de langues dans lesquelles une brève fermeture vélaire se rencontre à la fin des voyelles nasales, le mbay, le mixtec et le vietnamien. En bobo-fing, cette petite fermeture vélaire peut se réaliser comme une consonne épenthétique lorsqu'une voyelle suit le mot qui contient la voyelle nasale. Cette nasale épenthétique ne se rencontre qu'après les voyelles postérieures. (Nous notons le signe de la nasalité sous la voyelle, dans cet exemple et dans ceux qui suivent, afin de ne pas interférer avec les symboles qui marquent la tonalité).

tṵ̃ + ò > tṵ̃ŋ̀ò 'au marché'.

Du point de vue articulatoire, ce phénomène s'explique par le bref contact qu'il y a entre le dos de la langue et le voile du palais à la fin de l'articulation de la voyelle. L'examen de données de résonance magnétique (Demolin et Segebarth 1992; Demolin et al. 1995) montre clairement que pendant l'articulation des voyelles postérieures et de la voyelle ouverte [ã], il existe un contact entre la luette et le dos de le langue. On peut donc supposer que ce contact s'étend à la fin de la voyelle pour faire une brève occlusion qui donne une coloration vélaire à la voyelle nasale.
Dans le cas ou une postposition de type CV [ga] suit un mot qui contient une voyelle nasale, une nasale vélaire [ŋ] se réalise avant la consonne sonore [g].

zɔ̰̄ + gà > zɔ̰̄ŋgà 'il ne mange pas'.
sɔ̰̀ + gá > sɔ̰̀ŋá. 'ne pas cultiver'

Un cas similaire, entre le viel hindi et l'hindi moderne, est mentionné par Ohala et Ohala (1993: 238).

Remarquons cependant que devant une consonne sourde, le caractère vélaire de la nasale ne se manifeste pas.

tṵ̀ + tīmí > tṵ̀tīmí 'derrière le marché'.

Il existe encore un cas ou une nasale épenthétique peut se manifester, c'est celui où une consonne sonore autre que vélaire suit une voyelle nasale. Dans ce cas, une nasale homorganique de la consonne est insérée entre la voyelle et la consonne.

bó sɔ̰̀ + bé > bó sɔ̰̀mbé 'Cet homme là'
sḭ̀ + dǒ > sḭ̀ndǒ 'la bouche du soleil'

5. BIBLIOGRAPHIE

Beddor, P. S. 1982. *Phonological and phonetic effects of nasalization on vowel height.* Ph D dissertation, University of Minessota. Reproduced by Indiana University Linguistics Club.

Calliope. 1989. *La parole et son traitement automatique.* Paris Masson.

Chistovich, L. A. et V . V. Lublinskaya. 1979. The 'center of gravity' effect in vowel

spectra and critical distance between the formants. *Hearing Research* 1. 185-185.

Chistovich, L. A., R. L. Sheikin et V. V. Lublinskaya. 1979. 'Centers of gravity' and spectral peaks as the determinants of vowel quality. In B. Lindblom & S. Öhman (eds.). *Frontiers of speech communication research.* New-York. Academic Press. 143-157.

Demolin, D., J-M Hombert, V. Lecuit, A. Soquet et C. Segebarth. 1995. An MRI study of French vowels. *Eurospeech*, Madrid. 2235-2238.

Demolin, D. & Segebarth, C. 1992. Analyse de production de voyelles de quelques langues du Soudan central par IRM." in *Actes des XIXèmes Journées d'Etude de la Parole*, Université Libre de Bruxelles. 37-42.

Grégoire, C. & de Halleux, B. 1994. Etude lexicostatistique de quarante-trois langues et dialectes mande. " in *Africana Linguistica XI*, 142, Tervuren. 53-70.

Long, R.W. 1971. *A Comparative Study of the Northern Mande Languages*, Indiana, University Ph.D. Thesis.

Ohala, J. J. & Ohala, M. 1993. The phonetics of nasal phonology: theorems and data." in *Phonetics and Phonology, volume 5, Nasals, Nazalization and the Velum*, Academic Press. 225-249.

Ruhlen, M. 1978. Nasal vowels. Greenberg, J.(ed.), *Universals of Human Language*, Stanford, SUP. 202-241.

Welmers, W. 1958. "The Mande languages." in *Georgetown Series on Languages and Linguistics*, 11, p. 9-24.

COMPARAISON DES STRUCTURES SYLLABIQUES EN FRANÇAIS ET EN ANGLAIS

Jean-Philippe Goldman[1], Alain Content[1,2], Uli H.Frauenfelder[1]
goldman,content,uli@fapse.unige.ch
[1]Laboratoire de Psycholinguistique, Université de Genève, Suisse
tél: ++41.22.705.97.41 fax: ++41.22.300.14.82
[2]Laboratoire de Psychologie Expérimentale, Université libre de Bruxelles, Belgique

ABSTRACT

This study present a quantitative comparison between the phonological properties of French and English using the BRULEX and CELEX lexical databases, respectively. These lexical statistics give a description of the syllabic structure of these two languages. These analyses revealed some interesting differences in the size of the two respective syllabic inventories, in syllabic complexity and finally in the frequency distribution of the syllable types and tokens. The processing consequences of these differences are discussed.

1. INTRODUCTION

Dès les années 70, la difficulté à identifier des indices phonétiques invariants dans le signal de parole a conduit les psycholinguistes à envisager un rôle possible de la syllabe comme unité privilégiée d'accès au lexique. Cette hypothèse a été initialement renforcée par des travaux (Mehler et al., 1981) montrant la sensibilité des locuteurs francophones à la structure syllabique. Plus récemment, des comparaisons inter-langues ont conduit à l'hypothèse que l'importance de l'unité syllabique pourrait différer selon les propriétés phonétiques, phonologiques, prosodiques et distributionnelles des langues (Sebastiàn-Gallès et al., 1992). En particulier, les différences observées entre des locuteurs francophones et anglophones ont été expliquées en référence aux différences dans la complexité et le degré d'ambiguïté des structures syllabiques des deux langues (Cutler et al., 1986).

Cependant, les différences structurales entre les deux langues n'ont jamais été analysées en détail. Par ailleurs, les études sur la syllabation du français (Aubergé et al., 1988; Laporte 1988; Dell 1995) montrent des désaccords et des ambiguïtés non négligeables. C'est probablement un des facteurs qui expliquent qu'il n'existe que peu de données quantitatives sur l'inventaire et la distribution des syllabes du français (Wioland, 1995). Le travail présenté constitue une première tentative de préciser les différences structurales entre le français et l'anglais, par des analyses quantitatives sur des lexiques informatisés.

2. METHODE

De manière à pouvoir prendre en compte la fréquence d'usage des mots, nous avons sélectionné deux bases de données qui contiennent cette information, BRULEX (Content et al., 1990), pour le français (noté Fr), et CELEX (Celex, 1993), pour l'anglais (An).

BRULEX est basé sur le corpus du Micro-Robert et contient environ 35000 entrées. Après suppression des formes fléchies, des mots composés, des onomatopées, abréviations, et des emprunts, le corpus restant comporte 24494 entrées. L'information sur la fréquence d'usage est basée sur les données du Trésor de la Langue française. BRULEX ne fournit pas l'information sur la structure syllabique des représentations. Celle-ci a été générée sur base de l'analyse des groupes consonantiques intervocaliques dans le lexique BDLEX (Pérennou et al.).

La réduction du lexique anglais s'est faite en éliminant les mots composés, les abréviations et les mots empruntés, ramenant le corpus de 52447 à 40911 entrées. Un élément souvent considéré comme un facteur d'ambiguïté pour la syllabation de l'anglais est la notion d'ambisyllabicité, proposée notamment par Kahn (1980), selon laquelle une consonne intervocalique peut, dans certains contextes phonologiques, appartenir à la fois à la syllabe précédente et à la syllabe suivante. Pour évaluer les conséquences de ce facteur, nous avons introduit pour l'anglais une représentation re-syllabifiée (notée AnAmb).

Les conditions d'application de l'ambisyllabicité ne font pas l'objet d'un accord unanime. Nous avons ici opté pour l'analyse la plus conservatrice, c'est-à-dire celle qui donne lieu au moins de changements, sur base de l'algorithme proposé par Gussenhoven (1986). Il propose d'attacher à toute syllabe sans coda, la première consonne de l'attaque de la suivante à la condition que celle-ci comporte une voyelle réduite (par exemple, le mot 'villa', qui est codé phonétiquement vI.l@,

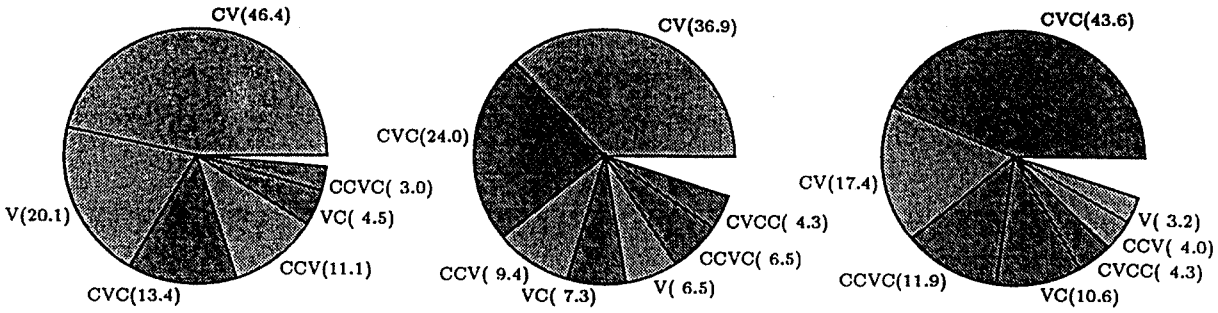


Figure 1: Répartition des structures syllabiques dans Fr., An., AnAmb. (pondérée par la fréq. d'utilisation)

devient vIl.l@, la consonne /l/ étant notée à la fois dans la première et la seconde syllabe).

Les analyses présentées portent sur la comparaison des structures et inventaires syllabiques des deux langues. Il faut remarquer que malgré la sélection effectuée, une grande différence persiste dans la taille des corpus.

3. RESULTATS

3.1. Inventaire phonémique et syllabique

Un premier indice éventuel de différences de complexité concerne le nombre absolu de syllabes et de structures syllabiques différentes attestées dans chacune des deux langues. La première analyse présentée concerne la taille de l'inventaire des segments et des syllabes. Les résultats principaux apparaissent dans la table 1.

Table 1: Inventaire phonémique et syllabique

	Fr	An	AnAmb
#mots	24.494	40.911	40.911
phonèmes	34	44	44
voyelles	15	23	23
consonnes	19	21	21
syllabes	3246	6781	8090
structures	21	22	22

Une différence importante apparaît en ce qui concerne l'inventaire des syllabes des deux langues. Le nombre de syllabes anglaises est plus de deux fois supérieur au répertoire syllabique français, et cette différence est encore renforcée par l'introduction de l'ambisyllabicité. Par contre, le nombre de structures syllabiques, définies en fonction de la classification des segments en consonnes (C) et voyelles (V) ne diffère pratiquement pas. En outre, le nombre de structures n'est pas modifié par la prise en compte de l'ambisyllabicité pour l'anglais.

Cette première analyse offre donc des indications contradictoires : la taille de l'inventaire syllabique confirme la plus grande complexité de l'anglais, mais il apparaît que la diversité des syllabes en anglais n'est pas liée à une plus grande variation de leur structure. L'analyse complète des relations entre ces deux observations dépasse le cadre du présent article, mais on peut d'ores et déjà signaler que la taille de l'inventaire phonémique (44 pour l'anglais vs. 34 pour le français) semble un facteur pertinent.

3.2. Fréquence des structures syllabiques

Une seconde analyse porte sur la fréquence des différentes structures syllabiques. On pourrait en effet faire l'hypothèse que malgré l'équivalence globale du nombre de structures, il existe des différences entre les deux langues en termes du nombre de structures syllabiques prédominantes. La figure 1 indique les proportions de syllabes des différents types dans chacun des corpus. Dans cette analyse, la fréquence d'usage des mots a été prise en compte, de telle sorte que la fréquence de chaque syllabe est pondérée par la fréquence d'usage des mots dans lesquels elle apparaît. Pour faciliter la lecture, les structures de fréquence inférieure à 3% ont été omises. Les résultats indiquent qu'un nombre très limité de structures syllabiques (6 ou 7) prédominent, et ils suggèrent effectivement des différences entre les deux langues: alors que 91.0% des syllabes attestées dans les mots français se répartissent entre 4 structures (CV, V, CVC et CCV), les quatre structures dominantes en anglais ne couvrent que 76.6% à 83.5%(pour An et AnAmb, respectivement). En outre, comme on peut le voir sur la figure 1, la proportion de syllabes non-couvertes par les 6 (Fr) ou 7(An) structures les plus fréquentes est plus importante en anglais.

La figure 1 montre également la nette prédominance des syllabes ouvertes (gris clair) par rapport aux syllabes fermées (gris foncé) en français. L'importance de la classe des syllabes ouvertes est fortement atténuée en anglais et la prédominance des syllabes fermées est encore renforcée lorsque l'ambisyllabicité est prise en compte (Cf. table 2). Il est intéressant de noter que, pour le français, nos observations s'avèrent très

proches de celles rapportées par Wioland, malgré les différences de procédure et de corpus. L'analyse réalisée par Wioland porte en effet sur un corpus de français parlé, transcrit et syllabé par l'auteur et il obtient un taux de 80.4% de syllabes ouvertes.

Table 2: Proportion de syllabes ouvertes

	Fr	An	AnAmb
Syll. ouv.	79.1%	54.1%	26.6%
Syll. fermées	20.9%	45.9%	73.4%
Ambiguïté	26.9%	39.6%	15.8%

On sait qu'en français, certaines voyelles apparaissent préférentiellement en syllabe ouverte et d'autres en syllabe fermée. Nous avons tenté d'estimer l'influence de ces contraintes phonotactiques en calculant la proportion de syllabes ouvertes et fermées pour chaque voyelle séparément. Le taux d'ambiguïté est la moyenne sur l'ensemble des voyelles de la plus petite des deux valeurs. Ainsi, un taux de 0 indiquerait que la voyelle n'apparaît que dans une classe de structure syllabique, tandis qu'un taux de 50% correspondrait au cas où les deux classes sont équiprobables. En moyenne, il apparaît que le taux d'ambiguïté est nettement plus élevé en anglais qu'en français. Par contre, cette différence est complètement éliminée si l'ambisyllabicité est prise en compte, du fait de la prédominance des syllabes fermées.

La différence dans le taux d'ambiguïté constitue donc un indice complémentaire en faveur de l'hypothèse que la structure syllabique peut être plus saillante en français. Contrairement à une suggestion assez répandue, il apparaît dans cette analyse que la prise en compte de l'ambisyllabicité pourrait simplifier, et non rendre plus ardue la segmentation syllabique, dans la mesure où, dans la plupart des cas, la voyelle attire la consonne subséquente.

3.3. Longueur des mots et des syllabes

Un indice potentiel de la complexité des syllabes est fourni par leur longueur en termes du nombre de phonèmes. La figure 2 (panneau gauche) montre la distribution des deux lexiques selon le nombre de phonèmes. Les longueurs moyennes sont similaires (6.61 et 6.74, resp. pour le français et l'anglais) et les distributions sont largement superposables. Mais, cette équivalence se réalise par des biais différents dans les deux langues. Comme l'indique le panneau médian, le mode du nombre de syllabes est de 3 pour le français et 2 pour l'anglais (moyenne resp.:2.83 et 2.62 syl./mot), et cette tendance est contrebalancée par la variation du nombre de phonèmes par syllabe (panneau droit, respectivement 2.33, 2.57 et 2.96 pour Fr, An et AnAmb en moyenne). Les syllabes comportant plus de trois phonèmes représentent une plus grande proportion pour le lexique anglais que pour le français. A l'inverse les syllabes formées d'un ou deux phonèmes sont en plus grand pourcentage en français. Cette relation d'échange correspond au principe général proposé par Menzerath (1954) qui stipule que le nombre de phonèmes par syllabe diminue à mesure qu'augmente le nombre de syllabes par mot. Nous avons donc examiné si la différence de longueur des syllabes entre les deux langues était observée indépendamment de la longueur des mots en syllabes. Les résultats (figure 3) montrent que si on considère la syllabation fournie par CELEX, les différences de longueur des syllabes s'estompent et disparaissent pour les mots longs par rapport au français. Par contre, pour la version ambisyllabifiée, la différence se maintient indépendamment de la longueur des mots.

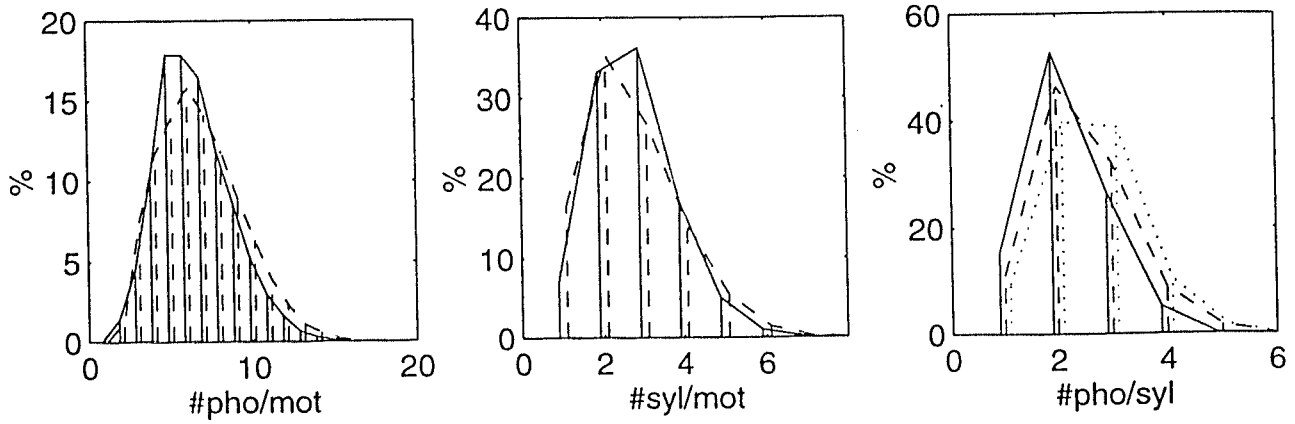


Figure 2: Distribution des mots et des syllabes selon la longueur (Fr: continu, An: tireté, AnAmb: pointillé)

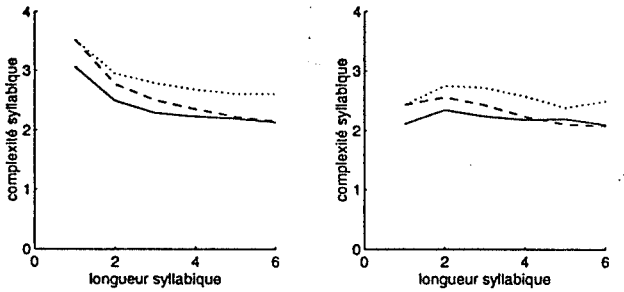


Figure 3: Longueur syllabique en fonction du nombre de syllabes par mot, sans (à gauche) et avec (à droite) pondération par la fréquence d'utilisation (Fr: continu, An: tireté, AnAmb: pointillé)

4. CONCLUSION

La comparaison quantitative des structures syllabiques en français et en anglais confirme l'existence de différences inter-langues. Nos analyses indiquent qu'en anglais, le nombre de syllabes attestées est plus grand, les structures syllabiques sont plus variées, les syllabes sont en moyenne plus longues. Toutefois, mis à part les variations de taille de l'inventaire, et la dominance des syllabes ouvertes en français, les différences observées ne sont pas extrêmement marquées. De plus,la prise en compte de l'ambisyllabicité pouvait dans le cas de l'ambiguïté, réduire l'écart entre les deux langues.

Le résultat le plus clair concerne la différence de taille des inventaires syllabiques des deux langues. La dimension du "lexique syllabique" pourrait effectivement constituer le point de départ d'une explication des différences dans le traitement de la chaîne parlée. Mais pour établir la nature de la relation entre les caractéristiques des deux langues et leur traitement, des analyses complémentaires plus fines selon la fréquence d'usage des mots seraient nécessaires, ainsi que l'examen de l'occupation de l'espace syllabique théorique des deux langues

En outre, une interprétation pertinente de ces données comparatives dépend aussi d'un modèle théorique du traitement plus explicite et de données expérimentales plus riches.

REMERCIEMENTS

Cette recherche a été financée par le F.N.R.S. 11-39553.93. Nous remercions F. Dell et P. Encrevé pour leurs commentaires et suggestions sur ces travaux.

5. BIBLIOGRAPHIE

Mehler J., Dommergues J.Y., Frauenfelder U., et Segui J. (1981) *The syllable role's in speech segmentation.* Journal of Verbal Learning and Verbal Behavior, 20:298-305.

Sebastiàn-Gallès N., Dupoux E., Segui J. et Mehler J.(1992) *Contrastive syllabic effect in catalan and spanish.*Journal of Memory and Language, 31:18-32.

Cutler A., Mehler J., Norris D. et Segui J. (1986) *The syllabe's differing role in the segmentation of french and english.* Journal of Memory and Language, 385-400.

Aubergé V., Boé L.-J., Lefèvre J.-P. (1988) *Lexiques et groupes consonantiques.*XVIIèmes J.E.P., 55-60

Laporte E. (1992) *Phonetic syllables in french: combinatorics, structure and formal definitions.* Acta Linguistica Hungarica, 41:175-189.

Dell F.*Consonant clusters and phonological syllables in french.* In press.

Wioland F. (1985) *Les structures syllabiques du français.* Slatkine.

Content A., Mousty P. et Radeau M. (1990) *Brulex. une base de données lexicales informatisée pour le français écrit et parlé.* Année Psycho., 90:551-556,

Celex (1993) *Centre for Lexical Information MPI for Psycholinguistics*, Nijmegen

Pérennou G. et de Calmès M. *BDLEX. Base de donnéees lexicales du français écrit et parlé.* Labo. CERFIA, Toulouse.

Kahn D. (1980) *Syllable-based generalizations in English phonology.* Garland, New-York.

Gussenhoven C. (1986) *English plosive allophones and ambisyllabicity* GRAMMA 10.2:119-141.

Menzerath (1954) *Die architektonik des deutschen wortschatzes.* Phonetische Studien,3

ÉTUDE ACOUSTIQUE DE VOYELLES DU FRANÇAIS EN CHANT ET EN PAROLE

Evelyne FLORIG

Institut de Phonétique — Université des Sciences Humaines de Strasbourg

22, rue Descartes — 67084 Strasbourg Cedex — e-mail : florig@ushs.u-strasbg.fr

ABSTRACT

The major question adressed in this study is, to find out to what extent the French vowels /i, e, ɛ, a, y/ undergo acoustic transformations as a fonction of frequency variation from soprano voices. The main trust of this research is to verify if the nature of spectral modification is comparable for all vowels.

1. INTRODUCTION

Par ses exigences spécifiques, par ses mouvements articulatoires complexes, le chant constitue un objet d'analyse particulier. En outre, en parole, l'objectif principal est la communication et donc la compréhension du message phonétique. En revanche, dans le chant, cette intelligibilité du message ne paraît pas constituer la qualité essentielle recherchée et semble passer au second plan. Par contre, la recherche de la justesse, l'aspect esthétique, la portée et la qualité de la voix jouent un rôle primordial et il s'agit de conserver une qualité de son homogène sur toute la tessiture. Chacun a déjà pu observer que la compréhension d'un texte chanté était plus problématique que celle d'un texte parlé et que, plus la voix devenait aiguë, plus la reconnaissance des phonèmes devenait difficile (cf. aussi J. Sundberg 1975, 1987, 1991 et B. Harmegnies & A. Landercy 1993).

Notre objectif a donc été de voir dans quelle mesure les voyelles /i, e, ɛ, a, y/ du français subissaient une transformation acoustique en fonction de la variation de la fréquence fondamentale.

2. DESCRIPTION DE L'EXPÉRIENCE

Nous avons enregistré deux sopranos lyriques du Conservatoire National de Région de Strasbourg à l'aide d'un appareil de type DAT.

2.1. Corpus

Comme le but de notre étude consistait en une comparaison acoustique de voyelles du français en parole et en chant, nous avons élaboré un corpus se composant de deux parties :

• une première partie en chant comportant des vocalises (figure 1) – en effet, la succession d'une même voyelle à des fréquences différentes remplit les conditions nécessaires à notre étude. Une série de vocalises ascendantes – avec hausse d'un demi-ton entre chaque vocalise – a été effectuée sur chacune des voyelles citées plus haut. Les vocalises sur la répétition d'une même voyelle à des fréquences différentes présentent l'avantage de ne pas comporter de consonnes intervocaliques, ce qui permet d'éviter les influences que ces consonnes auraient certainement eu sur le spectre acoustique des voyelles. En accord avec les chanteuses, nous avons fait débuter chacune des vocalises par le groupe consonantique /ts/, car il permettait d'évacuer le problème de l'attaque du son et semblait favoriser une bonne gestion du souffle pendant le chant (cf. aussi les remarques de R. Miller 1990 au sujet des consonnes initiales de vocalises).

• une seconde partie en parole : les vocalises étaient complétées par des phrases contenant les voyelles analysées. Nous les avons placées dans des mots, car il nous semblait plus naturel pour le locuteur de prononcer ces voyelles dans un contexte. Nous avons essayé de faire varier l'entourage de ces voyelles en les faisant précéder des consonnes / b, v, z, l, ʒ, ʁ / et en les plaçant en position accentuable. Nous avons choisi des consonnes orales, sonores, de différents lieux d'articulation et nous avons intentionnellement éliminé les consonnes sourdes et les consonnes nasales afin d'éviter une possible assimilation de sonorité ou de nasalité de la voyelle suivante.
Nous avions donc par exemple :
"Je vais chanter sur /i/, comme dans bise, vie, Suzie, lit, magie et Paris.
Je vais chanter sur /e/, comme dans bébé, privé, rusé, blé, léger et paré.
Je vais chanter sur /ɛ/, comme dans bête, vert,

zèbre, laide, geste et raide.
Je vais chanter sur /a/, comme dans bas, vase, visage, lame, jade et rat.
Je vais chanter sur /y/, comme dans butte, vue, zut, lutte, juste et rustre".

Les points de mesure ont été choisis à l'intérieur des voyelles afin d'éviter les influences des phonèmes voisins et les zones de transition formantiques.

2.2. Méthode d'exploitation

Nous avons numérisé les données acoustiques et nous les avons analysées informatiquement. Nous avons utilisé des spectrogrammes à filtrage large ou très large ainsi que des spectrogrammes à filtrage étroit.
Les mesures de fréquence fondamentale ont été effectuées sur des spectres à filtrage étroit et les mesures des formants ont été prises sur des spectres à filtrage large.

3. RÉSULTATS

3.1. Analyse en parole

Pour chacune de nos locutrices, nous avons effectué des mesures sur 31 /i/, 28 /e/, 17 /ɛ/, 24 /a/ et 17 /y/. Nous avons pris deux points de mesure par voyelle et nous avons établi des valeurs types pour chacune d'elles en faisant la moyenne des valeurs des formants que nous avons relevées (table 1).

Ces valeurs nous servent de point de référence dans notre analyse du chant, puisque que c'est par rapport au timbre des voyelles en parole que nous examinons dans quelle mesure la variation de fréquence a une influence sur la composition acoustique des voyelles.

3.2. Analyse en chant

Pour chaque locutrice, nous avons analysé 10 vocalises sur /i/, 8 sur /e/, 7 sur /ɛ/, 9 sur /a/ et 10 sur /y/. Comme chaque vocalise comporte 9 notes, nous avons analysé 90 notes pour /i/, 72 pour /e/, 63 pour /ɛ/, 81 pour /a/ et 90 pour /y/. Là aussi, nous avons pris deux points de mesure par note.

3.1.1 Établissement de seuils

Nous avons pu observer que la composition acoustique des différentes voyelles chantées n'était pas nette sur toute l'étendue de fréquence des vocalises. En effet, à partir de certaines notes aiguës, il devenait très difficile, voire impossible de distinguer les différents formants de manière précise. Nous avons pu remarquer que le premier harmonique et le fondamental étaient très rapidement supérieurs à la fréquence type du premier formant que nous avions établie pour la même voyelle en parole. C'est alors le Fo qui se trouvera renforcé et qui jouera le rôle de F1. Cependant, sa valeur atteint très vite un niveau trop élevé pour que les différents harmoniques puissent encore coïncider avec les zones de renforcement caractéristiques des voyelles. La structure formantique des voyelles n'a alors que très peu de ressemblance avec celle des voyelles de la parole qui nous servent de référence – voir par exemple la figure 2 ; il s'agit d'un spectre de /ɛ/ chantée sur do4, note sur laquelle la distinction des formants de cette voyelle devient problématique (cf. table 2). On peut observer ici que les zones de renforcement divergent de celles que nous avions relevées pour la même voyelle en parole (table 1).

À partir de valeurs moyennes, nous avons établi des seuils en-dessous desquels la structure formantique des voyelles est relativement proche de celle des mêmes voyelles en parole et au-dessus desquels, la localisation des différents formants devient problématique et ne correspond plus du tout aux valeurs de référence de la parole (cf. table 2) – N. Scotto Di Carlo (1972 et 1991) parle à ce propos de présence de "seuils d'intelligibilité", puisque l'intelligibilité d'une voyelle dépend de la netteté de ses formants (voir aussi G. Cornut & J.-C. Lafon 1960). Nous pouvons observer que les seuils que nous avons établis correspondent relativement bien avec ceux que signale N. Scotto Di Carlo : nous avons relevé que les voyelles /e/ et /ɛ/ avaient une composition acoustique assez proche de celle des mêmes voyelles parlées jusqu'à la3, note qu'elle considère comme limite d'intelligibilité correcte de ces voyelles, celles-ci étant altérées au-delà de ces hauteurs. De même, nous établissons un seuil à ré4 pour /a/ et N. Scotto Di Carlo fixe sa limite d'intelligibilité à mi4.

3.1.2. Variations de timbre

Nous avons comparé le timbre des voyelles chantées à celui des mêmes voyelles parlées en utilisant des moyennes (cf. figure 3). Nous avons établi 2 valeurs : la première dans une tessiture moyenne, relativement proche de la parole, et la deuxième dans une tessiture plus aiguë, correspondant au

deuxième seuil établi plus haut.

On peut observer que /i/ chanté est différent de /i/ parlé quelque soit la note émise : on observe une légère hausse de F1 – certainement due à l'action conjuguée de la hausse tonale et d'un abaissement du maxillaire – et une baisse de F2 – certainement due à une légère labialisation.

/e/ chanté est également caractérisé par une hausse de F1 et de F2 qui rejoint les valeurs de /i/ parlé.

/ɛ/ chanté a un timbre différent de/ɛ/ parlé dès les fréquences les plus graves puisque F1 est tout de suite supérieur et que F2 amorce d'emblée une baisse. Il semble que cette voyelles soit légèrement labialisée.

En chant, la voyelle /a/ a un F1 qui reste stable jusqu'aux alentours de ré4 et un F2 plus bas qu'en parole. Cette voyelle est vraisemblablement labialisée en chant.

Le F1 de /y/ chanté a tendance à augmenter – en raison, certainement, de l'abaissement du maxillaire accompagnant l'augmentation tonale – et le F2 amorce un légère baisse. Il nous semble que l'abaissement du maxillaire qui caractérise la montée dans les aigus entraîne une légère perte de labialité.

4. CONCLUSION

Nous avons pu observer que les variations de fréquence entraînent des modifications dans la compositions acoustique des voyelles. Il est possible de distinguer des seuils au-delà desquels la distinction des différents formants devient problématique. Nous avons pu voir, en outre, que le timbre des voyelles chantées est quelque peu différent de celui des mêmes voyelles parlées. Il s'agit d'une tendance générale se manifestant sur toutes les vocalises. Toutes les voyelles sont marquées d'une augmentation de F1, due, vraisemblablement, à l'action conjuguée de la hausse de fréquence et de l'abaissement du maxillaire inférieur. Les voyelles /i, ɛ, a/ semblent être légèrement labialisées tandis que la voyelle /y/ semble être caractérisée par une légère perte de labialité. Il serait utile de procéder à une étude perceptive afin d'expliquer pour quelles raisons, par quels mécanismes de compensation on perçoit toujours des voyelles alors qu'on ne distingue plus les formants. Par ailleurs, il nous semble intéressant, voire essentiel, de tenter de mettre en évidence les modifications articulatoires intervenant dans le chant. Nous commençons ainsi une étude articulatoire à partir de films radiologiques du conduit buccal et de films vidéo des mouvements labiaux qui nous aideront, espérons-le, à comprendre le mécanisme de compensations mis en oeuvre par les chanteurs.

5. BIBLIOGRAPHIE

Cornut G., Lafon J.-C. (1960), Étude acoustique comparative des phonèmes vocaliques de la voix parlée et chantée, *Folia Phoniatrica*, vol. 12, n°3, pp. 188-196.

Harmegnies B., Landercy A. (1993) Analyse spectrale et voix chantée. Contribution à une métrologie objective de la maîtrise du chant, *Revue de phonétique appliquée,* n°106, pp. 35-59.

Miller R. (1990) *La structure du chant*, trad. de Gouëlou J.-M., Paris : IPMC.

Scotto Di Carlo N. (1972) Étude acoustique et auditive des facteurs d'intelligibilité de la voix chantée, *Travaux de l'Institut de Phonétique d'Aix*, vol.1, pp. 115-129.

Scotto Di Carlo N. (1991) La voix chantée, *La Recherche*, XXIII, 235, pp. 1016-1025.

Sundberg J. (1975) Formant Technique in a Professional Female Singer, *Acustica*, 32, 2, pp. 89-96

Sundberg J. (1987) *The science of the singing voice*, Illinois : Northem Illinois University Press.

Sundberg J. (1991) Phonatory vibrations in singer's. A critical Review, *STL-QPSR*, vol.1, Stockholm.

Sundberg J., Johansson C., Wilbrand H. (1982) X-Ray Study of Articulation and Formant Frequencies in two female Singers, *STL-QPSR*, vol.4, Stockholm.

Table 1. Valeurs de référence des formants en paroles – notées en Hertz.

	F1	F2	F3
/i/	334	2627	3295
/e/	378	2317	3136
/ɛ/	635	2109	3080
/a/	732	1727	2980
/y/	329	2021	2688

Tableau 2. Seuils de fréquence

	composition acoustique distincte	composition acoustique imprécise
/i/	jusqu'à la#3	de la#3 à mi4
/e/	jusqu'à sol#3-la3	de la#3 à ré4
/ɛ/	jusqu'à la3	de la3 à do#4
/a/	jusqu'à do#4-ré4	de do#4-ré4 à mi4
/y/	jusqu'à si3-do4	de si3-do4 à ré4-ré#4

Figure 1. Exemple de vocalise

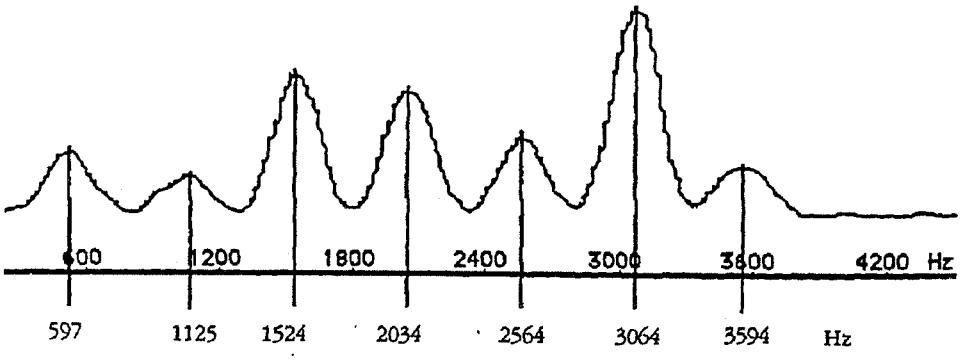


Figure 2. Spectre de /ɛ/ chanté sur do4.

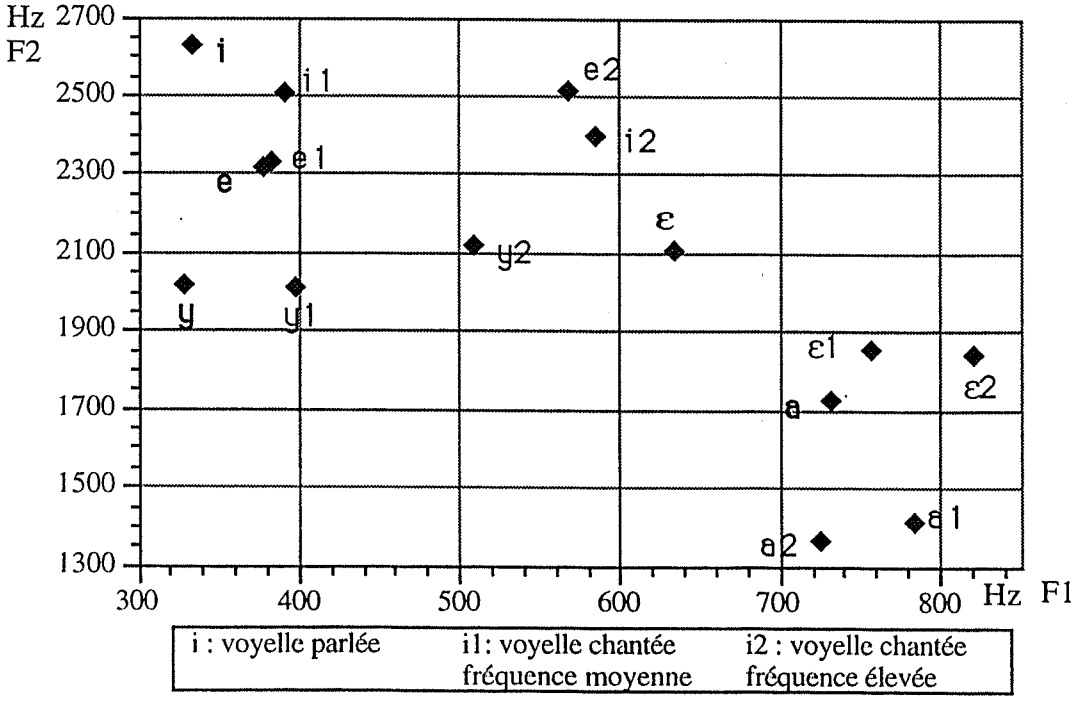


Figure 3. Valeurs formantiques des voyelles parlées et chantées

ETUDE COMPARATIVE DES VOYELLES AMERICAINES /i I e/ - L'INFLUENCE DE L1 SUR L2

Diana TREVINO-SIGMUND

ILPGA et UA1027, CNRS-19, rue des Bernardins, 75005 Paris
e-mail : trevino@msh-paris.fr

ABSTRACT

This paper is a contribution to the research dealing with the influence of a learner's native language (L1) in the production and perception of a foreign language (L2). The production of the front american vowels /i I e/ by french speakers at a slow and fast rate is compared to that of native american speakers. The parameters, duration, Fo, and formants constitute the basis of the analysis. The results confirm that the phonemes of L2 which are not existant in the learner's native language are assimilated to the closest phoneme in the learner's L1. The results concerning F2 of /i/ are not in agreement with the literature.

1. INTRODUCTION

L'intérêt porté à l'apprentissage d'une langue seconde a encouragé les études comparatives sur la production et la perception de la parole. Des travaux traitant de la perception chez le nourisson et ses capacitiés perceptives (*language universal*) (Aslin *et al.*, 1981 ; Jusczyk, 1984 ; Werker *et al.*, 1984a) et de leur diminution au contact de la langue de l'environnement (*language specific*) (Werker *et al.*, 1984a, 1988 ; Best, 1991) figurent également dans ce champ. L'expérience linguistique semble avoir une influence sur la perception de la parole. Cette adaptation aurait des conséquences dans l'apprentissage d'une langue seconde à l'âge adulte. Les tests de perception se limitent cependant à quelques contrastes (essentiellement le lieu d'articulation dentale, labiale et vélaire et le voisement), et les conclusions ne sont pas toujours évidentes. Les études de Pisoni *et al.* (1982) montrent qu'avec de l'entraînement, les apprenants peuvent distinguer des paires de phonèmes n'existant pas dans leur langue maternelle. Les tests de perception sont également mis en cause. Les tests d'identification, par exemple, sont donnés parfois aux sujets qui ne connaissent ni les sons ni les symboles phonétiques. Nous nous intéressons ici à des constrastes fins entre les voyelles antérieures fermées.

Un des problèmes rencontrés par les français dans l'apprentissage de l'anglais américain est la distinction entre les voyelles /i/, /I/, et /e/. D'après Delattre (1965), les voyelles américaines tendent à être diphtonguées et changent de timbre au cours du temps ; la différence de timbre ne serait pas la seule en cause : les voyelles tendues sont plus longues que les relâchées (Lehiste, 1960). Les voyelles françaises sont plus stables et plus périphériques que les voyelles américaines (Delattre, 1965).

Notre travail est donc une étude qui met en évidence le filtre phonologique de L1 dans la production et perception des voyelles américaines /i I e / par des sujets français.

2. PROCEDURE

6 sujets masculins, 3 Parisiens et 3 Américains ont participé à notre étude. Les 3 sujets français avaient plus ou moins 4 années d'études dans L2. Les sujets américains suivaient des cours de français à Paris. Les deux groupes de participants avaient entre 25 et 45 ans. Les voyelles américaines ont été intégrées dan la phrase cadre : "He said the word feed (CVC) dabbler" (Il a dit le mot /fid/, /fId/, /fed/ bricoleur). Les phrases ont été répétées 3 fois par les locuteurs français et américains en débit lent et rapide. En débit lent Fo et les formants ont été mesurés à 30ms après l'"onset" de la voyelle, au centre et à 30ms avant la fin du signal acoustique de la voyelle. En débit rapide, nous avons mesuré à 10ms après l'"onset" de la voyelle, au centre et à 10ms avant la fin du signal sonore de la voyelle.

3. RESULTATS

Les figures 1, 2, 3, et 4 concernent les voyelles /i I e/ pour les locuteurs français et américains pour les deux débits.

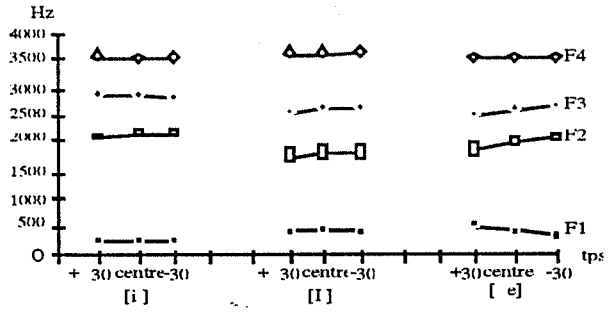


Figure 1 : Valeurs formantiques moyennes pour [i I e] réalisées par des locuteurs américains en débit lent.

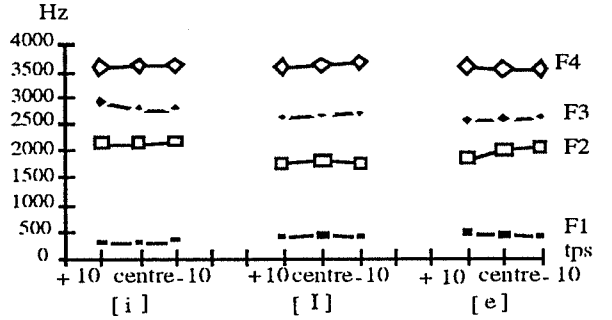


Figure 2 : Valeurs formantiques moyennes pour [i I e] réalisées par des locuteurs américains en débit rapide.

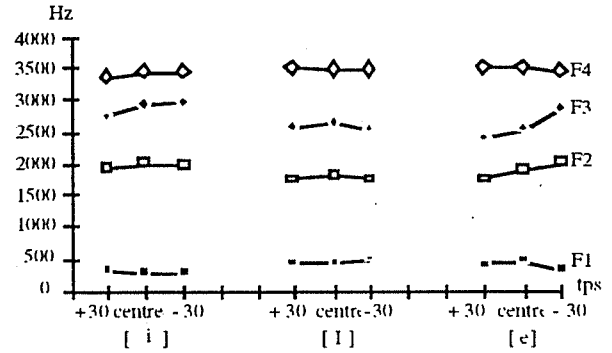


Figure 3 : Valeurs formantiques moyennes pour [i I e] réalisées par des locuteurs français en débit lent.

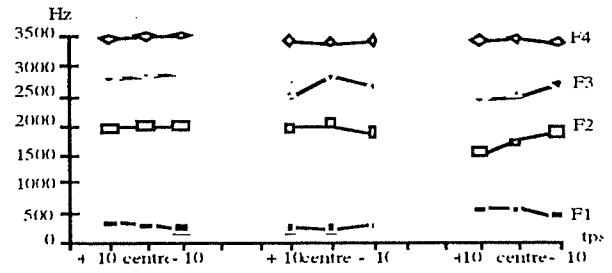


Figure 4 : Valeurs formantiques moyennes pour [i I e] réalisées par des locuteurs français en débit rapide.

Il y a peu de différence de formants entre les deux débits pour les locuteurs américains (figures 1 et 2). En débit lent F3 et F4 de /i/ sont plus proches qu'en débit rapide. En débit rapide, à 10ms après l'"onset" de la voyelle, F3 descend vers la cible. Nous interprétons cette chute de F3 comme un arrondissement des lèvres plus important en débit rapide qu'en débit lent. Dans les deux débits, F2 de /i/ est supérieur à 2kHz. Pour /i/ et /I/, F2 est stable et le "undershoot" (Lindblom *et al.*, 1994) n'est pas évident. Cela est sans doute dû au fait que la consonne labiodentale précédant les voyelles n'influence pas le mouvement de la langue. Pour la voyelle /e/ en débit lent (figure 1) la diphtongaison est très nette : F1 descend doucement, F2 monte ("glide") suivi par une zone plus ou moins stable, et F3 accuse une montée importante, signalant un écartement des lèvres (Selon Lehiste et Peterson, les voyelles cibles sont caractérisées par des zones stables et lorsqu'il y a un mouvement d'un ou deux formants intervenant entre les zones stables on trouve un "glide" (Lehiste, 1960). Par contre, pour /e/ en débit rapide (figure 2), la pente de F1 est plus douce qu'en débit lent et celle de F3 est moins accusée, car les lèvres sont moins étirées.

Pour les 3 locuteurs français, dans les deux débits (figures 3 et 4), F3 et F4 de /i/ sont proches et F2 en dessous de 2kHz. Les lèvres seraient maximalement étirées (F3 élevé) mais la langue dans une position antérieure non extrême (F2 non maximalement élevé). En débit rapide (figure 4), on peut interpréter la voyelle /I/ relâchée comme la réalisation de deux vocoïdes (un vocoïde étant un son vocalique n'appartenant pas à un système phonologique). Pour la voyelle /e/ avec un mouvement important de F3 (figure 4), nous avons constaté la réalisation de deux vocoïdes, la première ressemble à la voyelle /e/ française, et la deuxième à la voyelle /i/. Alors que chez les américains, il y un "glide" plus une partie stable.

En ce qui concerne la durée en débit lent (figure 5), nos résultats suggèrent qu'il n'y a pas une différence très nette entre la moyenne de durée de /i/ tendu et /I/ relâché aussi bien chez les locuteurs américains que chez les locuteurs français, contrairement à ce qui est souvent affirmé dans la littérature (Lehiste, 1967). La compensation (Lindblom, 1979; Maeda, 1989), l'économie des gestes (Martinet, 1955 ; Lindblom, 1986), et des variations entre locuteurs, pourraient expliquer ces différences.

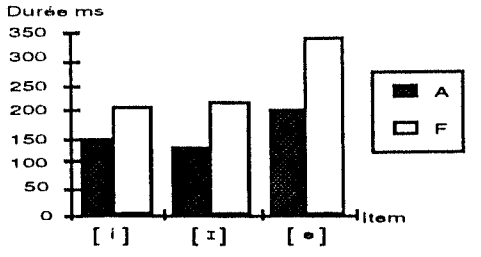


Figure 5 : Moyennes des durées des voyelles américaines /i I e/ réalisées par des locuteurs américains et français en débit lent.

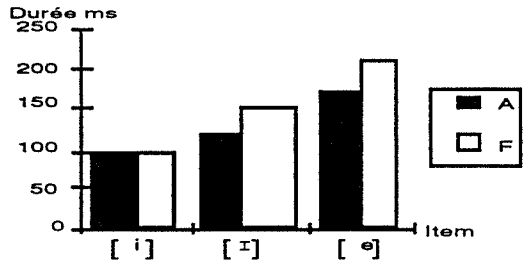


Figure 6 : Moyennes des durée des voyelles américaines /i I e/ réalisées par des locuteurs américains et français en débit rapide.

Les résultats de la figure 5 suggèrent que la moyenne de durée de ces mêmes voyelles est, chez les locuteurs français, supérieure à celle des locuteurs américains. En ce qui concerne la voyelle /e/ diphtonguée, la moyenne de durée française est beaucoup plus longue que la moyenne américaine. Nous pensons que cette différence est probablement due à des facteurs suivants : 1) la réalisation de deux vocoïdes (voir figures 3 et 4), 2) la tendance à parler plus lentement en L2, et 3) la volonté de réaliser 2 cibles successivement en lieu d'un /e/ diphtongué car il n'y a pas de diphtongues en français.

Quant à la durée en débit rapide (figure 6), notre graphique suggère que la voyelle /I/ est plus longue que la voyelle /i/ chez les deux groupes de locuteurs. Les voyelles /I/ et /e/ sont plus longues chez les locuteurs français que chez les locuteurs américains.

4. TEST DE PERCEPTION

Nous avons fait passer un test de discrimination à 10 étudiants de phonétique de Paris III. Le test portait sur deux groupes de stimuli, l'un sur la durée et l'autre sur les timbres des voyelles.

Les voyelles /i I e ɛ/ dans le contexte /mVd/ ont été artificiellement soutenues pendant une longue durée. Puis nous avons extrait 350ms au cœur de chaque voyelle. Pour la voyelle /e/ nous avons inclus la partie finale de la voyelle contenant le "glide". La durée du deuxième stimilus était de 250ms. Elles ont été présentées en paire de 350ms et 250ms dans un ordre aléatoire (et à intervalle de 4s).

Pour la partie portant sur la durée, 4 auditeurs sur 10 ont confondu les différences de durée pour la voyelle /ɛ/.

Pour la partie portant sur le timbre, nous avons utilisé les paires de voyelles suivantes : /e-ɛ/, /i-I/, /ei-e/, /I-i/, /i-ɛ/, /e-ɛ/, /I- i/, /ei-e/, /i-I/, /e-i /.

Les résultats montrent que la plupart des auditeurs ont éprouvé des problèmes dans la deuxième partie du test. 6 auditeurs sur 10 ont confondu la paire /ei-e/ et 5 auditeurs le paire /i-I/.

5. CONCLUSION

Les conclusions sont les suivantes : 1) Pour la voyelle /i/ F3 est plus élevé pour les Français que pour les Américains. 2) F2 de /i/ est plus haut pour les américains que pour les Français, contrairement à la littérature. Il y des explications possibles : a) Dues aux locuteurs : il y a deux façons de réaliser un /i/ en français, soit de jouer sur F3 (F3 élevé), soit d'antérioriser la langue (Schwartz, *et al.*, 1993 cf. Modélisation de Maeda, 1989). b) Ces différences sont probablement dues à des raisons d'ordre contextuel. Le locus de la consonne /d/ française aurait une influence sur F2. Donc F2 pourrait être moins élevé chez les locuteurs français.

Pour la voyelle américaine /I/ relâchée et la voyelle /e/ diphtonguée les locuteurs français ont réalisé 2 vocoïdes semblables aux voyelles françaises / e/ et /i/.

Quant à la durée, aucun groupe de locuteurs n'a fait de distinction de durée entre les voyelles tendues et relâchées. Nous savons que si l'on raccourci la voyelle /i/ à 10ms près, on entend la voyelle /I/. La perception de /i/ n'étant pas symétrique pourrait influencer la perception de /i/ donc, la durée. En débit lent, la durée de la voyelle /e/ diphtonguée était beaucoup plus longue chez les locuteurs français que chez les locuteurs américains et ce qui est probablement dû à la réalisations de deux vocoïdes. Selon Heffner (1964), les mesures de durée ne tiennent pas compte de la coarticulation. D'après lui le rapport entre les données acoustiques et la coarticulation n'est pas une chose simple à définir. Par le test de perception, nous avons constaté, que le timbre joue un rôle plus important que la durée, au moins dans les limites de notre expérience.

6. BIBLIOGRAPHIE

Aslin, R. N., Pisoni, D. B., Hennessy, B. L & Perey A. J., (1981). Discriminaition of onset time by human infants : new findings and implications for the effect of early experience. Child Development, 52, 1135-1145.

Best, C. T. (1991). Phonetic influence on the perception of non-native speech contrasts by 6-8 and 10-12 months olds. In the *Biennial meeting of the Society for Research in Child Development*, Seattle, WA.

Delattre, P. (1965) *Comparing the Phonetic Features of English, French, German, and Spanish*, Julius Verlag, Heidelberg.

Heffner, R-M. S. (1964) *General Phonetics*. The

university of Wisconsin Press, Madison, Wis.
Jusczyk, P. W. (1984) On characterizing the development of speech perception. In J. Mehler, E. C. Walker & M. Garrett (Eds.), *Neonate cognition: beyond the blooming, buzzing confusion*, 199-211, Hillsdale, NJ : Erlbaum.
Lehiste, I., Peterson, G. E. , (1960) *Readings In Acoustic Phonetics*, The MIT Press, Ma. 1967.
Lindblom, B. (1986) Economy of speech gestures. In *Speech Production* (P. F. MacNeilage, editor) 217-245, New York : Springer-Verlag.
Lindblom, B., Moon, S-J., (1994) Interaction between Duration, Context, and Speaking Style in Stressed vowels, *JASA* n° 96 : 40-55.
Lindblom, B., Sundberg, J., (1979) Acoustical consequences of Lip, Tongue, Jaw, and Larynx Movement, *JASA* n° 50 : 1166-1179.
Maeda, S. (1989) Articulation compensatoire des voyelles : analyse de données cinéradiographiques avec un modèle linéaire. *Mélange de Phonétique générale et expérimentale offerts à Pela Simon.* 545- 62, Strasbourg : Institut de Phonétique.
Maeda, S. Programme de simulation acoustique du conduit vocal (VTCALC), 1989.
Martinet, A. (1955). *La linguistique synchronique*, PUF, Paris.
Pisoni, D. B., Aslin, R. N. Perey, A. J., Hennessy, B. L. (1982) Some effects of laboratory training on identification and discrimination of voicing contrast in stop consonants, *Journal of Experimental Psychology : Human Perception and Performance*, n° 8, 297-314.
Schwartz, J-L., Beautemps, D., Abry, C., Escudier, P., (1993) Inter-individual and cross-linguistic strategies for the production of the [i] vs. [y] contrast, *Journal of Phonetics* n° 21, 411-425.
Werker, J. F., Lalonde, C. E. (1988) Cross-language speech perception : Initial capabilities and developmental change, *Developmental psychology* n° 24(5), 672-683.
Werker, J. F., & Tees, R. C. (1984a) Cross-language speech perception : Evidence for perceptual reorganization during the first year of life, *Infant Behavior and Development*, n° 7, 49-63.

ÉTUDE DES PHÉNOMÈNES DE RÉDUCTION VOCALIQUE EN ANGLAIS BRITANNIQUE

Gaëlle FERRÉ

ILPGA - 19, rue des bernardins - 75005 Paris

Tél.: (1) 44 32 05 75 - Fax: 43 29 70 13

ABSTRACT

Within pairs of words such as "ex'hibit, exhi'bition", it has already been proved that shifting the accent caused a centralisation of the vowel (Delattre, 1969), and the longer the word is, the shorter the duration of the stressed vowel. We wanted to check this hypothesis, and to know whether there really is a centralisation of the vowel (reduced vowels would then merge towards [ə]), or a reduction of the formant target (reduced vowels would then be a "weak" version of full vowels). It appeared that stress is not correlated to the syllable intensity, as well as to subglottal pressure. As far as the rate of speech is concerned, we found that vowels are centralised when they are pronounced at a fast rate of speech.

1. INTRODUCTION

Dans la parole continue, par rapport aux mots isolés, il y a trois stratégies possibles pour le locuteur:

- STRATÉGIE 1: Priorité au contraste consonnes-voyelles: F1 a tendance à être plus élevé pour toutes les voyelles, quelque soit le contexte dans lequel elles se trouvent: le locuteur ouvre plus la bouche, sans doute pour augmenter le contraste entre voyelles et consonnes successives.

- STRATÉGIE 2: Priorité aux consonnes: F1 est plus bas pour toutes les voyelles: ce sont les consonnes qui ont un effet d'assimilation d'ouverture (les consonnes ont intrinsèquement un F1 plus bas que les voyelles).

- STRATÉGIE 3: Priorité aux voyelles: F1 est plus bas pour les voyelles fermées et plus haut pour les voyelles ouvertes: les traits d'aperture sont alors accentués.

En premier lieu, nous analyserons donc le déplacement de l'accent comme la première cause de réduction vocalique en anglais, puis l'effet du débit de parole.

2. DÉPLACEMENT DE L'ACCENT

Trois locuteurs britanniques de sexe masculin (de 40 à 50 ans), à Londres, ont lu 6 paires de mots, situées dans la phrase cadre "Say the word... again". Les mots contenaient les voyelles cibles antérieures, postérieures et centrales suivantes:

	acc.	inacc.	ANT	POST
i	com'peting [kɒm'pitɪŋ]	compe'tition [kɒmpə'tiʃn]	+	-
I	ex'hibit [Ig'zIbIt]	exhi'bition [eksI'bIʃn]	+	-
æ	a'dapting [ə'dæptɪŋ]	adap'tation [ədə'teiʃn]	+	-
ɜ	con'versing [kɒn'vɜsɪŋ]	conver'sation [kɒnvə'seiʃn]	-	-
ʌ	in'substance [In'sʌbstæns]	insub'stantial [InsʌbstænʃI]	-	+
U	you 'should be [jU 'ʃUd bI]	you sho u l d 'stay [jU ʃUd 'stei]	-	+

Matériel utilisé pour les mesures: UNICE et Signalyze.

Nous avons mesuré F0, F1 et F2, ainsi que leur amplitude A1 et A2 au début, au centre et à la fin de la voyelle. Nous avons donc obtenu pour chaque paramètre, et par locuteur, 18 mesures pour les voyelles accentuées et 18 mesures pour les voyelles inaccentuées, soit au total 54 mesures pour les voyelles accentuées et 54 mesures pour les voyelles inaccentuées, par paramètre, si l'on regroupe les mesures des trois locuteurs.

F0 baisse de 18,7 % entre voyelles accentuées et voyelles inaccentuées; F1 baisse de 6,6 %, F2, de 6,5 %. A1 baisse de 28,1 %, et A2 de 35,4 %. C'est la durée qui est le plus affectée par le déplacement de l'accent, car elle baisse de 50 % entre voyelles accentuées et voyelles inaccentuées. La durée des voyelles inaccentuées est en moyenne deux fois plus courte que celle des voyelles accentuées (table 1).

Le test statistique (ANOVA) montre que le déplacement de l'accent a un effet significatif

sur la durée et sur l'amplitude des deux premiers formants. En revanche, elle n'a pas d'effet significatif sur F0, F1 et F2 (table 1).

Dans le cas d'une réduction causée par le déplacement de l'accent, il n'y a pas centralisation de la voyelle réduite, mais celle-ci n'atteint pas sa cible vocalique; si elle avait atteint sa cible vocalique, elle aurait du être identique à la voyelle qui se trouvait en position accentuée, dans chaque paire de mots (figure 1). (STRATÉGIE 2: translation du triangle vocalique vers les basses fréquences).

2. 1. Effet du déplacement de l'accent sur l'intensité et la pression sous-glottique

Un locuteur britannique (45 ans environ) a lu les séries de mots suivantes deux fois.

competing, **sport** , **super**
superior, **pic**ture, **par**tition
pattern, **expect**, re**touch**
polygamist, **poo**dle, com**pe**tition
support, **superb**, **superiority**
picturesque, **par**titive, **para**chute
expectant, **touch**y , **poly**glot
pouch, com**pe**te, **support**
superbly , **pic**torial , **para**cetamol
expectation , **touch**ingly , **poly**gon
port , **expect**ing, re**port**age
report, **two**, **push**

Chaque mot étant lu deux fois, nous avons donc pris deux mesures d'intensité et de pression sous-glottique par mot, à l'aide de l'aérophone II[1] . Certaines mesures n'ont pu être reportées, soit que les courbes d'intensité et de pression sous-glottique étaient anormalement basses, soit que la courbe du volume d'air ne revenait pas à zéro entre les mots. Nous avons donc effectué un test de corrélation sur 58 mesures d'intensité et 58 mesures de pression sous-glottique, que nous avons corrélées avec l'accentuation, et nous avons trouvé que le déplacement de l'accent n'a pas d'effet significatif sur la pression sous-glottique puisque p = 0,23, ainsi que sur l'intensité, où p = 0,091 (table 2).

Cependant, on constate que l'intensité est en moyenne plus élevée sur les syllabes accentuées que sur les syllabes inaccentuées (on obtient une moyenne de 76,5 dB pour les syllabes accentuées contre 71 dB pour les syllabes inaccentuées). De même, la pression sous-glottique est en moyenne plus élevée sur les syllabes accentuées que sur les syllabes inaccentuées (avec 12,5 cm H2O pour les syllabes accentuées et 11,3 cm H2O sur les syllabes inaccentuées) (Table 2).

3. LE DÉBIT

3. 1. Le débit et l'accentuation

Les mêmes locuteurs que lors de l'expérience 1 ont lu une phrase en adoptant différents débits de parole, dans les mêmes conditions:

Trois types de débits étudiés: normal, rapide et lent. (Pour le débit rapide, nous avons demandé à nos locuteurs de lire la phrase aussi rapidement que possible, et pour le débit lent, de lire cette phrase très lentement).

Corpus: He had much diffidence of his own capacity, except in a field of battle.

[hɪ həd ˈmʌʧ ˈdɪfɪdəns əv hɪz ˈəʊn kəˈpæsɪtɪ əkˈsɛpt ɪn ə ˈfiəld əv ˈbætl]

Pour nos mesures, nous n'avons pas pris en compte la dernière syllabe de "diffidence", ainsi que la syllabe "own", à cause de la nasale.

Nous avons mesuré, avec le même matériel, F0, F1, F2, A1, A2, au centre de chaque voyelle, en débit normal, rapide et lent, pour chaque locuteur. Ceci nous fait un total de 18 mesures par débit, et par paramètre étudié pour chaque locuteur, avec 6 voyelles accentuées et 12 voyelles inaccentuées. Nous avons également calculé le débit moyen, ainsi que le rapport entre débit normal et rapide, et entre le débit normal et lent (table 3).

Lorsque l'on fait varier le débit dans une phrase, il apparaît que l'accentuation a un effet significatif sur tous les paramètres (ANOVA).

F0	p = 0,0007	A1	p = 0,0015
F1	p = 0,0001	A2	p = 0,0001
F2	p = 0,0009	Durée	p = 0,0001

3. 2. Effet du débit sur les paramètres segmentaux et suprasegmentaux sans prendre en compte l'accentuation

Le débit n'a pas eu d'effet significatif sur F0. Il a un effet significatif sur F1, mais pas sur F2 (plus grande ouverture du conduit vocal en débit lent, d'où une augmentation du

[1] Cette expérience a été réalisée au laboratoire de l'hôpital Laënnec, PARIS, grâce au Dr. S. Hans, que nous tenons à remercier ici, ainsi que le responsable du laboratoire, le Dr. Brasnu.

contraste consonnes/ voyelles). Il n'a pas d'effet significatif sur A1 et A2. La durée des voyelles est plus longue en débit lent, puis normal, et enfin rapide. Les voyelles des mots grammaticaux sont aussi plus brèves que celles des mots lexicaux. (Table 4). Il serait intéressant de calculer, dans une prochaine étude, le rapport entre les voyelles et les consonnes, afin de savoir si ce rapport change lorsque l'on fait varier le débit.

En ce qui concerne le mode de réduction vocalique, on constate qu'il y a centralisation des voyelles (figure 2).

4. CONCLUSION

Cette étude montre qu'en anglais, les voyelles se distinguent moins bien et sont plus obscures lorsqu'elles se trouvent en position inaccentuée que lorsqu'elles occupent une position accentuée. La raison, nous l'avons montré, est que la voyelle est physiquement et physiologiquement "réduite", et l'IPA emploie d'ailleurs des symboles phonétiques différents pour noter cette réduction, comme c'est le cas de "competing" [kɒm'pitIŋ], "competition" [kɒmpə'tiʃn]. Cependant, l'IPA emploie les mêmes symboles phonétiques pour le [I] de "exhibit" et de "exhibition", ainsi que pour le [æ] de "adapting" et "adaptation", alors qu'il y a également réduction de la voyelle. La réduction ne s'opère cependant pas de la même façon en anglais britannique et en anglais américain.

Par comparaison avec l'anglais américain (cf. D. Trevino-Sigmund, 1994), on a trouvé qu'en anglais britannique, les voyelles réduites ne sont pas centralisées lorsque la réduction est due à un déplacement de l'accent, et que la phrase est prononcée à un débit normal. En revanche, les deux premiers formants des voyelles sont plus bas, ce qui revient à dire que la cible vocalique n'est pas atteinte, et que dans ce cas, la voyelle ne constitue qu'un "passage articulatoire" d'une consonne à l'autre. Par contre, en débit rapide, nous avons trouvé qu'il y a centralisation des voyelles. C'est donc dans ce cas que les voyelles sont réduites de la même façon en anglais britannique et en anglais américain.

En ce qui concerne l'accentuation, nous pouvons dire que celle-ci est corrélée avec la durée: les voyelles accentuées sont plus longues de 50% que les voyelles inaccentuées. Elle est également corrélée avec l'amplitude de F1 et F2, qui est plus élevée pour les voyelles accentuées que pour les voyelles inaccentuées (28,1% et 35,4%). En revanche, nous n'avons pas pu établir de corrélation entre l'accentuation et la courbe du fondamental, de même qu'entre l'accentuation et les deux premiers formants. Nous avons également trouvé que le débit ne fait pas varier de manière significative la courbe du fondamental, F2, non plus que l'amplitude de F1 et F2. Par contre, il joue un rôle important quant à la durée des voyelles, qui est plus longue en débit lent et plus courte en débit rapide, et également en ce qui concerne F1.

Enfin, il aurait été intéressant de parler du style d'articulation, et de comparer les résultats avec ceux obtenus avec des débits de parole différents, mais cette étude fera l'objet d'un prochain article.

5. FIGURES ET TABLES

Table 1: Moyennes de F0, F1, F2, A1, A2 et de durée (36 mesures de durée) pour les voyelles accentuées et inaccentuées des trois locuteurs. (Moyennage sur toute la voyelle réalisé sur 54 mesures pour chaque paramètre). Différences et pourcentage de baisse entre voyelles accentuées et inaccentuées pour tous les paramètres, ainsi que les résultats du test ANOVA.

	F0 (Hz)	F1 (Hz)	F2 (Hz)	A1 (dB)	A2 (dB)	Durée (ms)
Acc	139	392	1627	32	31	84
Inacc	113	366	1520	23	20	42
Diff.	26	26	107	9	11	42
% de baisse	18,7	6,6	6,5	28,1	35,4	50
p	0,08	0,57	0,38	0,024	0,02	0,0003

Table 2: Moyennes de pression sous-glottique (H2O) et d'intensité (dB) pour les syllabes accentuées et les syllabes inaccentuées, réalisées sur 58 mesures de pression sous-glottique et 58 mesures d'intensité, et effet de l'accentuation sur les deux paramètres (ANOVA).

	pression ss-glottiq. (H2O)	Intensité (dB)
Acc.	12,5	76,5
Inacc.	11,3	71
p	0,23	0,091

Table 3: Moyennes des temps de parole (ms), temps de pause (ms), débit (rapport de proportionnalité entre débit rapide et normal, et entre débit lent et normal), temps d'articulation (ms), durée totale des voyelles étudiées (ms), et durée totale des consonnes (ms), en débit normal, rapide et lent.

	normal	rapide	lent
tps de parole (ms)	4220	2910	6383
pause (ms)	395	*	715
débit	*	+ 1,3	- 1,5

tps d'art. (ms)	3825	2910	5668
voyelles (ms)	1647	1370	1962
consonnes (ms)	2178	1540	3706

Table 4: Moyennes des mesures de F0, F1, F2, A1, A2 et de durée, obtenues pour les trois locuteurs, et effet du débit sur tous les paramètres (ANOVA).

	normal	rapide	lent	p
F0 (Hz)	135	126	128	0,26
F1 (Hz)	397	363	442	0,0005
F2 (Hz)	1594	1567	1616	0,49
A1 (dB)	28,8	29	27	0,16
A2 (dB)	26,2	25	26,3	0,88
Durée (ms)	101	82	114	0,0001

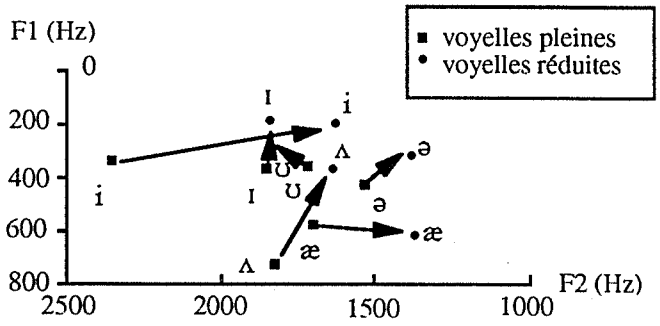


Figure 1: Réduction vocalique due au déplacement de l'accent (moyenne des résultats obtenus pour les trois locuteurs).

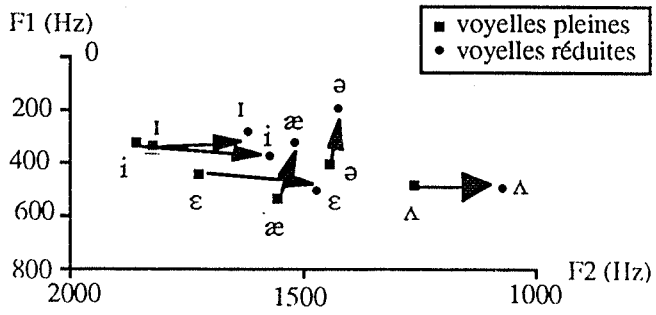


Figure 2: Réduction des voyelles en débit rapide, par rapport au débit normal (moyenne des résultats obtenus pour les trois locuteurs).

6. BIBLIOGRAPHIE

Delattre, P. (1968). From Acoustic Cues to Distinctive Features, *Phonetica*, n° 18, 198-230.

Delattre, P. (1969). An acoustic and articulatory study of vowel reduction in four languages. *IRAL*, n° 8, 295-325.

Fant, G. (1960). *Acoustic Theory of Speech Production*. The Hague, Mouton.

Frokjaer-Jensen, B. (1992). Data on Air Pressure, Mean Flow Rate, Glottal Input and Output, Energy, Aerodynamic Resistence, and Glottal Efficiency for Normal and Healthy Voices. A Preliminary Study. *22ème congrès mondial de l'IALP*, Hanover.

Ladefoged, P. (1962). ELEMENTS OF ACOUSTIC PHONETICS. Chicago, University of Chicago Press.

Ladefoged, P. (1975). *A Course in Phonetics*. New York, Harcourt Brace, Jovanovitch.

Lindblom, B. (1963). Spectrographic Study of Vowel Reduction. *JASA*, n° 35, 1773-1781.

Lindblom, B. (1963). *On Vowel Reduction*. Speech Transmission Laboratory, Stockholm.

Manuel, S. Y. (1991). Recovery of Deleted Scwha. *PERILUS IV: Current Phonetic Research Paradigms: Implications for Speech Motor Control*, Stockholm.

Manuel, S. (1992). Vowel Reduction and Percetual Recovery in Casual Speech. *JASA*, n° 91(4), 2388- ?.

Trevino-Sigmund, D. (1994). *Etude comparative entre les voyelles anglo-américaines /i I e/ et les voyelles françaises /i e/.* Mémoire de DEA. Paris III.

Van Bergem, D. R. (1995 a). *Acoustic and Lexical Vowel Reduction*. PhD, University of Amsterdam, Studies in Language and Language Use, Vol.16.

Van Bergem, D. R. (1995 b). Reflections on Aspects of Vowel Reduction. *IFA Proceedings*, University of Amsterdam 18.

GROUPES CONSONANTIQUES ET ÉPENTHÈSE EN TURC

Aline ASCI

Institut de Phonétique - Université des Sciences Humaines de Strasbourg - 22, rue Descartes - 67084 Strbg. Cedex
Tél. : 88 41 73 69 - Fax : 88 41 73 69 - e-mail : asci@ushs.u-strasbg.fr

ABSTRACT

The aim of this investigation is to examine the appearence of "intrusive" vowels in consonant clusters for Turkish.

It is demonstrated that such epenthetic vowels emerge between clusters that present different places of articulation. It thus follows that such vowels would not appear when consonants have similar places of articulation. Articulatory and acoustic data indicate the presence and robustness of this vocalic element.

1. INTRODUCTION

En turc, le nombre de groupes consonantiques est très limité. Les emprunts possédant un groupe de consonnes sont adaptés aux contraintes phonotactiques de la langue (Deny, 1955) par l'adjonction d'une voyelle prothétique comme -i- dans [istasjɔn] (station), [istatisik] (statistique) ou épenthétique comme -u- dans [gurup] (groupe) ou encore -i- dans [filim] (film).

Selon Carton (1974) l'épenthèse vocalique ou la voyelle d'appui est définie comme "***un son vocalique que l'on ajoute à l'intérieur d'un groupe de consonnes, ce qui rend l'articulation plus conforme aux habitudes de la langue. On l'appelle aussi parfois svarabhaktique*** ".

L'apparition de cet élément épenthétique semble être beaucoup plus proéminente lorsqu'il s'agit d'emprunts que lorsqu'on a affaire aux groupes consonantiques intrinsèques au turc. Cependant, nos analyses oscillographiques ont mis en évidence l'existence de cette voyelle d'appui même lorsqu'il s'agit de ces groupes consonantiques du turc.

Cet élément dont la présence semble être latente dans la conscience phonologique du locuteur, est bien présent, intercalé entre deux consonnes.

Le but de cette étude est de montrer l'apparition de cette voyelle épenthétique et d'examiner ses contextes d'apparition.

2. MÉTHODE

2.1. Sujets

2 locuteurs turcs, un homme et une femme, tous deux âgés de vingt-cinq ans, ont servi de sujets. Ils présentent des caractéristiques assez homogènes et leur prononciation n'est teintée d'aucun particularisme régional.

2.2. Corpus

Le corpus comprend soixante phrases regroupant un maximum de groupes consonantiques.

2.3. Enregistrement

Nos documents d'analyse sont des oscillogrammes (Brock, 1989), fournissant des informations sur quatre paramètres articulatori-acoustiques : laryngal, nasal, résonance buccale et pression d'air à la sortie buccale. Cette méthode nous donne aussi des renseignements sur les battements de l'apico-alvéolaire [r].

3. RÉSULTATS ET DISCUSSION

L'analyse des groupes consonantiques avec une liquide nous a permis de constater le fait suivant : tantôt entre la liquide [l, r] et la consonne suivante un élément épenthétique apparaissait, tantôt cet élément était absent.

Une investigation minutieuse, nous a permis de constater que ce phénomène survenait lorsque la liquide était suivie d'une consonne occlusive. Par exemple /arka/.

L'un des groupes consonantiques possibles en turc est la suite: **liquides + occlusives**, (Ergin, 1962), suite, elle-même subdivisée en trois catégories :

-liquides + alvéodentales [, r + t, d]
-liquides + vélaires [l, r + k, g]
-liquides + bilabiale (nasale: m) [l, r + m]

3.1. Réalisations d'épenthèses

Nous avons constaté deux types de réalisations différentes avec les liquides [l, r] suivis de consonnes occlusives.

3.1.1. Liquides + occlusives alvéodentales

Dans les groupes de consonnes de type [l, r] + alvéodentales, notre analyse s'est portée sur les exemples suivants :

orta	[orta]	"moyen"
arda	[arda]	"derrière"
alti	[altɯ]	"six"
ürdün	[yrdyn]	"la Jordanie"
örnek	[ørnɛk]	"exemple"
erdi	[ɛrdi]	"il est arrivé à"
geldi	[gɛldi]	"il est venu"

Lors de la réalisation de ce type de groupe consonantique, notre analyse a mis en évidence l'absence d'élément d'appui entre les deux consonnes. En effet, les consonnes impliquées dans ce type de groupe de consonnes avaient leur lieu d'articulation relativement proches.

En effet, [r, l] sont des constrictives apico-alvéolaires, [t, d, n] sont des occlusives alvéodentales. Lors du passage de [r] ou de [l/]à [t, d, n], nous ne remarquons aucun changement de position de la langue : il en résulte une réalisation du groupe consonantique sans l'appui d'élément épenthétique.

3.1.2. Liquides + vélaires

En ce qui concerne la catégorie [l] et [r] suivis des occlusives vélaires [k] ou [g], notre analyse a pris en compte les exemples suivants :

erkek	[ɛrkɛk]	"homme"
türkü	[tyrky]	"chanson"
tilki	[tilki]	"renard"
ilgi	[ilgi]	"intérêt"
karga	[karga]	"corbeau"

Dans ces groupes consonantiques, nous décelons la présence d'un élément vocalique [ə]. L'apparition de cet élément est provoquée par la différence des lieux d'articulation de ces deux sortes de consonnes. [k] et [g] sont des occlusives dorso-vélaires, alors que [l] et [r] sont des constrictives apico-alvéolaires. Ainsi le changement de la position de la langue d'avant vers l'arrière s'accompagne dc la réalisation de l'élément épenthétique [ə], qu'on peut considérer comme une voyelle d'appui.

Dans les tableaux ci-dessous, nous présentons les durées, en *cs*, de l'élément épenthétique [ə] qui s'intercale entre la liquide et les consonnes occlusives vélaires, chez nos deux locuteurs.

Tableau 1: Résultats pour le sujet 1

	Sujet 1		
	liquide	ə	occ.vél.
ɛrkɛk	5	3	13
tilki	4	4	15
tyrky	4	4	15
ilgi	5	2	10
karga	4	5	9

Tableau 2: Résultats pour le sujet 2

	Sujet 2		
	liquide	ə	occ.vél.
ɛrkɛk	3	3	10
tilki	4	3	10
tyrky	4	3	10
ilgi	4	2	7
karga	4	3	6

Ces mesures, ainsi que les données obtenues pour les autres paramètres articulatori-acoustiques, nous permettent de constater que les valeurs de la durée de la voyelle d'appui [ə] sont aussi importantes que celles des liquides [l] ou [r].

En comparant les durées de [ə] avec celles des sons de son entourage, nous pensons que [ə] peut être considéré comme une voyelle à part entière qui possède une durée significative, même si elle n'apparaît pas dans la graphie.

3.1.3. Liquide + bilabiale /m/

La réalisation du groupe consonantique /rm/ dans les exemples tels que [ɯrmak] (rivière) et [orman] (forêt) s'effectue également avec l'insertion de la voyelle épenthétique [ə] entre les deux consonnes, voyelle qui correspond à un changement de lieux d'articulation lors du passage de la consonne [r] à l'occlusive [m].

Les durées, en *cs*, de la voyelle d'appui [ə] présente dans ce groupe consonantique, figurent dans les tableau suivant :

Tableau 3: Résultats pour le sujet 1

	Sujet 1		
	liquide	ə	m
orman	4	4	9
ɯrmk	5	5	8

Tableau 4. Résultats pour le sujet 2

	Sujet 2		
	liquide	ə	m
orman	3	4	10
urmk	4	3	7

Pour cette catégorie de groupes de consonnes également, les durées le l'élément épenthétique [ə] sont équivalentes ou même supérieures aux durées de la consonne [r] placée devant l'occlusive bilabiale nasale [m].

Si l'on compare les durées de la voyelle d'appui [ə] figurant dans les quatre tableaux ci-dessus avec les durées des liquides [l, r], nous observons que [ə] possède des durées égales et même parfois supérieures aux durées des liquides.

La moyenne de durée de l'élément d'appui [ə] est de 3,5 *cs* pour S1 et de 3 *cs* pour S2.

Les valeurs relevées dans nos analyses se rapprochent de celles de Quilis (1970) pour l'espagnol qui affirme que "la durée de l'élément vocalique est très variable, allant de 0,8 *cs* à 5,6 *cs*. La moyenne de sa durée est de 3,2 *cs*." Il ajoute aussi que "cet élément vocalique possède une structure acoustique très semblable à celle d'une voyelle".

Toutes ces données renforcent notre conviction que cette voyelle d'appui est une voyelle à part entière et qu'elle a une existence effective sur le plan phonétique.

Sachant que le système orthographique du turc est plutôt phono-logique, en d'autres termes, à chaque son correspond une représentation graphique, pourquoi cette voyelle d'appui n'apparaît - t- elle pas dans la graphie ?

Pour répondre à ces questions nous proposons deux tentatives d'explication.

1. D'une part, sur le plan phonétique, nous avons constaté que cet élément épenthétique n'apparaissait que dans des contextes où les groupes de consonnes avaient leurs lieux d'articulation distincts. En dehors de ce contexte, il n'y a pas d'émergence de [ə].

La présence de cette voyelle d'appui est ainsi strictement d'ordre phonétique : elle n'est pas prise en compte par le système qui met sur le même plan les groupes de consonnes partageant le même structure au plan phonologique.

Il en va de même en français où l'apparition d'une voyelle épenthétique dans les exemples comme "jardin", "bras", "serment"...n'est pas non plus prise en compte par le système de la langue.

2. D'autre part, si la langue turque ne tient pas compte de l'existence de cet élément dans la graphie, c'est parce que du point de vue de la structure syllabique, il n'y a pas de création de syllabe supplémentaire dans ces cas-là. En effet, sachant que la coupe syllabique s'effectue entre les deux consonnes formant les groupe consonantique (V+Liquide / Occ.+V), nous pouvons conclure que [ə] n'est pas considéré comme noyau de syllabe.

4. CONCLUSION

Pour faciliter la prononciation de mots empruntés contenant un groupe consonantique, le turc fait appel à une voyelle épenthétique qui s'intercale entre les deux consonnes de ce groupe, comme par exemple [filim] ou [gurup]. Mais il existe aussi un certain nombre de groupes de consonnes intrinsèques à la langue turque où, dans la graphie, cette voyelle épenthétique n'apparaît pas.

Or, comme nous avons constaté dans nos recherches, l'absence orthographique de cette voyelle d'appui n'exclut pas son existence sur le plan phonétique.

Les analyses s'appliquant aux groupes de consonnes intrinsèques à la langue turque, nous permettent d'aboutir aux conclusions suivantes :

- dès lors que dans un groupe consonantique les lieux d'articulation de ces consonnes sont proches voire identiques, il n'y a aucun élément épenthétique inséré entre ces deux consonnes.

- par contre, si les lieux d'articulation de ces consonnes sont distincts, un élément épenthétique [ə] s'intercale entre ces deux consonnes.

En outre, les mesures des durées nous ont permis de voir que la durée de l'élément épenthétique [ə] était égale ou supérieure aux durées des liquides [l, r].

Toutes ces données nous amènent à penser, comme Basbøll (1988), que les groupes consonantiques devraient être resyllabifiés en accord avec les contraintes phonotactiques de la langue en question. "It is widely agreed that

clusters should be syllabified in agreement with the phonotactics of the language in question." Cela est d'autant plus vrai pour le turc qui a une structure syllabique -(C)VC- au même titre que le japonais.

5. BIBLIOGRAPHIE

Basbøll (1988) *Phonological theory in Linguistic theory foundations*. Cambridge University Press, Cambridge.

Brock G. (1989) *Optimalisation d'une station analogique de traitement du signal acoustique. In Mélanges de Phonétique Générale et Expérimentale offerts à Péla SIMON*. Publications de Insitut de Phonétique de Strasbourg.

Carton F. (1974) *Introduction à la phonétique du français*. Bordas.

Deny J. (1955) *Principes de Grammaire Turque*. Adrien Maisonneuve, Paris.

Ergin M. (1962) *Türk Dil Bilgisi*. Minnetoglu Yayinlari, Istanbul..

Quilis A. (1970) *El elemento esvarabatico en los grupos [pr, br, tr...]*. In Mélanges offerts à G. Straka, Tome1, Lyon-Strasbourg.

6. DOCUMENTS

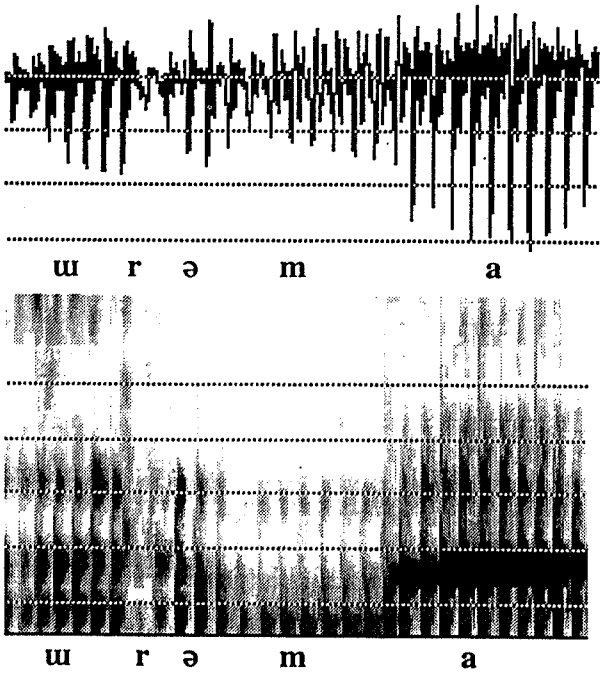


Figure 1 : Oscillogramme et spectrogramme synchrones pour la séquence [ɯrmak], où entre [r] et [m] apparaît la voyelle d'appui [ə].

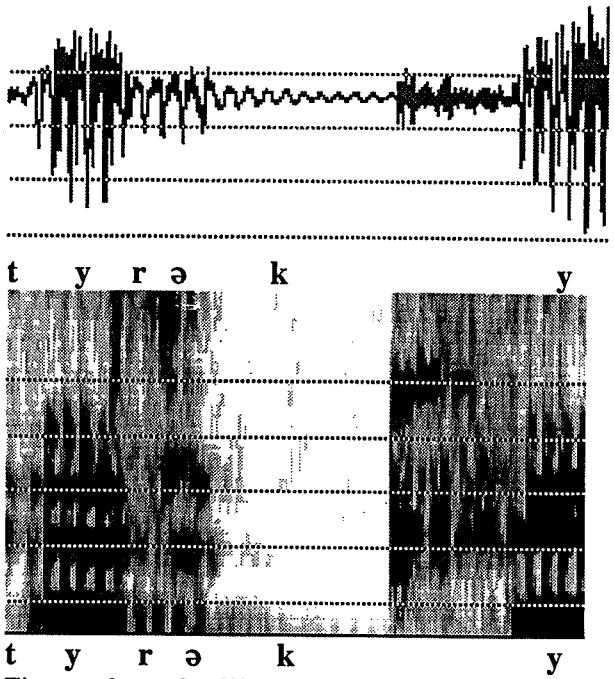


Figure 2 : Oscillogramme et spectrogramme synchrones pour le séquence [tyrky], où entre [r] et [k] apparaît la voyelle d'appui [ə].

La matérialité des structures sonores du langage
1. Taxinomies phonologiques et tendances universelles

Nathalie Vallée, Louis-Jean Boë, Christian Abry, Jean-Luc Schwartz, Ahmed Berrah

ICP UPRES A 5009 INPG/Université Stendhal, BP 25 38040 Grenoble Cedex 9
Tél : 76 82 43 38 – Fax : 76 82 43 35 – e-mail : (vallee, boe, abry, schwartz, berrah)@icp.grenet.fr

ABSTRACT

Are the main tendencies of phonological systems of the world languages constrained by production and perception? This question is discussed within the framework of the "substance–based" linguistic approach suggested simultaneously by Lindblom and Stevens in 1972. We present in this paper several universal tendencies of phonological systems which could be explained by materiality of sound structures and we discuss them in the light of ontogenic data.

1. TAXINOMIES PHONOLOGIQUES

La taxinomie, qui vise à proposer un ordre à l'intérieur duquel sont classés des éléments d'un ensemble, constitue un des programmes fondateurs de toute science : Où commence la variété ? Quelles sont les limites d'une espèce ? Sur quels critères doit-on rassembler les éléments d'une même famille ? Au-delà de la complexité et de la diversité des éléments, les procédures taxinomiques s'efforcent de trouver les ressemblances – les traces d'une même organisation –, de maîtriser les différences, afin de détacher les caractères extérieurs pour pouvoir accéder à des structures de moins en moins factuelles et de plus en plus générales.

La démarche taxinomique en linguistique est en pleine maturation mais elle est encore loin du stade qu'avaient atteint la botanique et la zoologie au XVIII^{e} siècle. La tâche n'est pas aisée. Ainsi on recense dans le monde de 2000 à 8000 idiomes selon les auteurs, tout simplement parce qu'il n'existe pas de consensus sur la méthode permettant de tracer les limites entre langue et dialecte, ni sur les critères permettant d'opérer les classifications. Dans l'état actuel des recherches, on est encore loin de disposer d'un échantillon représentatif des descriptions linguistiques pour les langues du monde (si la notion d'échantillon a ici un sens), alors que nombre d'entre elles sont en voie de disparition rapide (Ladefoged, 1995).

C'est Troubetzkoy (1939) qui a véritablement inauguré l'ère des taxinomies phonologiques et des premières recherches sur les universaux :

> « J'ai mis au net tous les systèmes vocaliques que je connaissais par cœur (34 en tout) et j'ai essayé de les comparer les uns aux autres [...]. Les résultats sont extrêmement curieux... Tous les systèmes se réduisent à un petit nombre de types et peuvent toujours être représentés par des schémas symétriques... Plusieurs lois « de la formation des systèmes » se laissent dégager sans peine... » (1928).

De telles recherches vont se poursuivre en relation avec les études sur l'ontogenèse du langage (Jakobson, 1941) et la définition des traits phonétiques (Hockett, 1955).

Par la suite, les travaux les plus marquants vont s'effectuer aux USA dans une démarche descriptive associée à une recherche sur les universaux (Greenberg, Ferguson et Moravcsik, 1978). Après le travail d'archivage, publié depuis 1944 dans l'*International Journal of American Linguistics*, et celui des missionnaires formés par le *Summer Institute of Linguistics*, cette quête est marquée par : 1° les travaux de Greenberg dans les années 1950 et la *Conference on Language Universals* à New-York en 1961 ; 2° la discussion des propositions de Chomsky au *Symposium on Universals in Linguistic Theory* à Austin, en 1967. *The Language Universal Project* (1967-1976) a permis l'élaboration et la diffusion des *Stanford Phonology Archives* et des propositions de typologie (Ruhlen, 1987). Les tendances universelles de systèmes phonologiques vocaliques alors présentées par Sedlak (1969) portent sur 150 langues, celles de Crothers (1978) reposent sur 209 descriptions.

S'inscrit dans ce courant de recherche la base de donnée UPSID (*UCLA Phonological Segment Inventory Database*) élaborée par Maddieson (*Patterns of Sounds,* 1984). Elle contient 317 systèmes, la dernière version a été étendue à 451 (Maddieson & Precoda, 1989). Rappelons que le choix des 20 familles retenues pour l'échantillonnage s'appuie sur la classification de Stanford : khoisan, niger-kordofanien, nilo-saharien, afro-asiatique, dravidien, burushaski, caucasien, indo-européen, basque, ouralo-altaïque, aïnou, paléo-sibérien, esquimau-aléoute, sino-tibétain, austro-thaï, austro-asiatique, indo-pacifique, australien, nord et sud américidien.

2. DES UNIVERSAUX VOCALIQUES ET CONSONANTIQUES

UPSID représente un précieux matériau pour la mise en évidence des typologies et la prédiction des structures sonores. Implantée à l'ICP, UPSID$_{317}$ a été extrêmement utilisée pour des recherches typologiques sur les voyelles, les

diphtongues et les consonnes. Sur le même modèle d'implantation a été constituée la base de données régionale RHÔNSON pour 43 parlers de la zone franco-provençale, avec des représentants des zones française, occitane et alémanique adjacentes (Jomaa & Abry, 1992). Nous présentons ici seulement les résultats typologiques sur les structures vocaliques et consonantiques qui vont servir à notre propos (Vallée, 1989, 1994 ; Vallée & al., 1990 ; Bernatova & Zackova, 1992 ; Berrah, 1994 ; Schwartz & al., 1996).

2.1. Les voyelles

Les systèmes recrutent 3 à 24 voyelles. Les tendances universelles déterminées d'après $UPSID_{317}$ sont présentées dans la table 1b, avec en ligne le nombre de voyelles de base et en colonne les systèmes (les plus fréquents à gauche, et à droite, ceux qui arrivent en seconde position) ; pour chaque système figure le pourcentage d'occurrence. Le système à 5 /i 'e' a 'o' u/ est de loin le mieux représenté. C'est le cas dans 3 des 4 grands groupes linguistiques (eurasien, américain, australasien), alors que c'est le système à 7 /i e ɛ a o ɔ u/ qui est majoritaire en Afrique (Maddieson, 1991). Pour les systèmes pairs à 6 et 8 voyelles, les plus fréquents possèdent une voyelle centrale de type /ə/ qui pourrait renvoyer à une autre dimension (Schwartz & al., 1996). Au-delà de 9 timbres de base les langues utilisent une autre dimension, la nasalité en général, ou la quantité. Cette relation tendancielle a déjà été proposée par Vallée (1989, 94) qui a quantifié la relation entre nombre de voyelles et timbres de base.

2.2. Le rapport entre les voyelles et les consonnes

Pour l'ensemble des langues la corrélation entre le nombre de consonnes et le nombre de voyelles n'est pas significative. Seule régularité, chaque langue possède plus de consonnes que de voyelles (à 2 exceptions près : le pawaian avec 12 voyelles et 10 consonnes ; et l'apinaye, 17 voyelles et 13 consonnes) (figure 1).

Table 1a : Les voyelles de la table 1b.

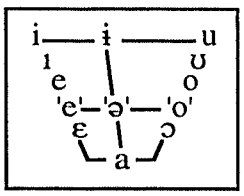
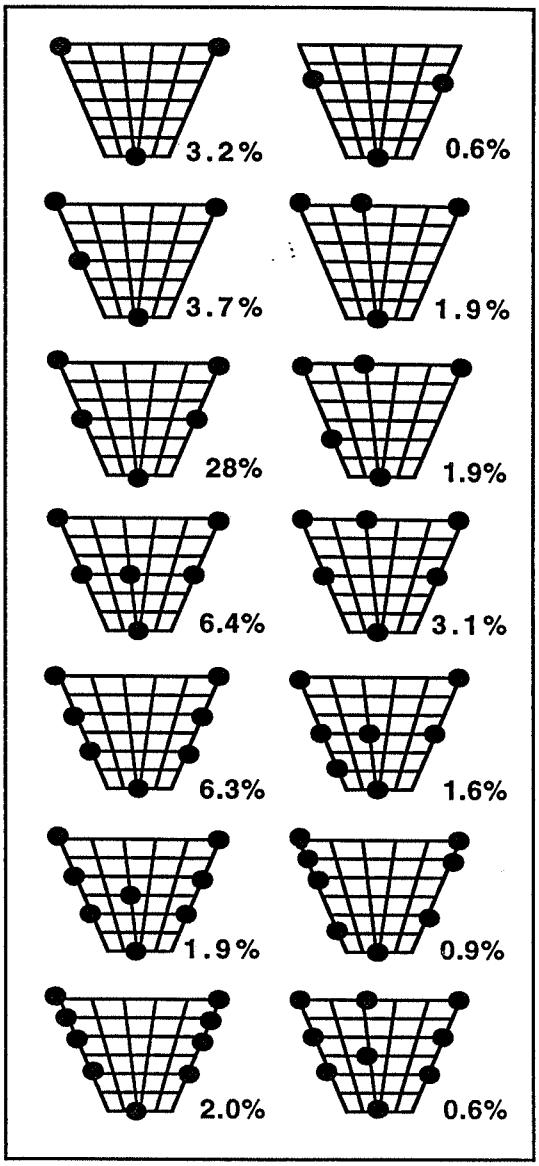


Table 1b : Tendances universelles des systèmes vocaliques d'$UPSID_{317}$ (d'après Vallée, 1994).



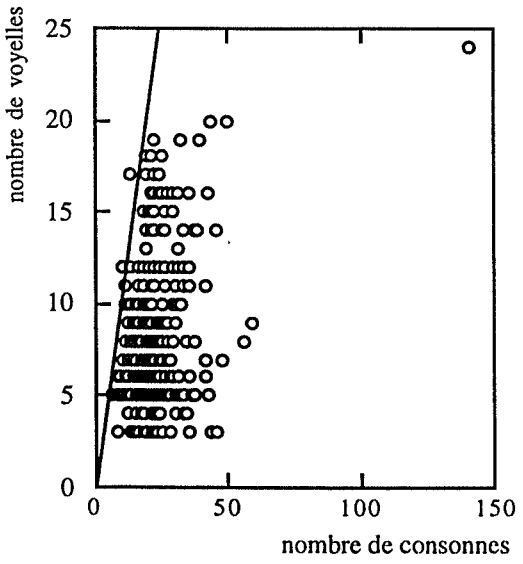


Figure 1 : Nombre de voyelles en fonction du nombre de consonnes dans les systèmes d'$UPSID_{317}$.

2.3. Les consonnes

Les systèmes phonologiques présentent le plus fréquemment 18 à 21 consonnes : min 6 pour le rotokas et max 141 (dont 48 clicks) pour le !xũ. Elles se répartissent ainsi par mode : occlusives 38.6%, fricatives 20.2%, nasales 14.6%, approximantes 13.0%, mi-occlusives 9.6%, vibrantes ou battues 3.9% ($UPSID_{451}$).

Les sourdes sont majoritaires dans les catégories mi-occlusives (74%), fricatives (72%) et occlusives orales (64%). Pour chaque langue, les occlusives (orales + nasales) sont beaucoup plus nombreuses que les fricatives (figure 2).

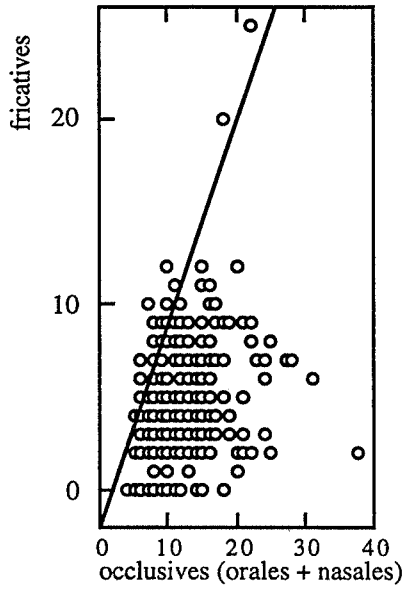


Figure 2 : Nombre de fricatives en fonction du nombre d'occlusives (orales+nasales) dans $UPSID_{317}$.

Si l'on comptabilise dans $UPSID_{451}$ la distribution des occlusives par lieu d'articulation, trois classes majeures sont quasiment présentes dans tous les systèmes : coronales 100% (dentales, alvéolaires, alvéo-dentales et post-alvéolaires), bilabiales 99% et vélaires 99% ; alors que : palatales 17% et uvulaires 14%. Mais on note le bon score des glottales avec 46%. Les occlusives pharyngales, qui n'étaient pas répertoriées dans $UPSID_{317}$, apparaissent dans 0.7% des langues.

La table 2 présente, par ordre décroissant, les pourcentages d'occurrences des 25 consonnes les plus fréquentes relevées dans $UPSID_{451}$.

3. UNIVERSAUX ET ONTOGENÈSE

Chez les enfants de 11 à 12 mois (au stade du babillage canonique ou panaché), la revue de question de Locke (1983), comptant 129 sujets pour 15 langues, fait apparaître l'inventaire suivant, soit dans l'ordre décroissant : /b/ et /m/, /p/, /d/, /h/ et /n/, /t/, /g/ et /k/, /j/ et /w/, /s/ ; ces consonnes rendant compte de 95% des occurrences. On ne dispose pas d'un tel inventaire aussi précis pour les voyelles, mais elles sont beaucoup moins nombreuses que les consonnes.

Ces 12 consonnes se retrouvent exactement – au /l/ près qui est acquis plus tardivement – parmi les consonnes les plus fréquentes dans l'inventaire des langues de la base, tel que cela apparaît dans la table 2.

Dans ce peloton de tête les *occlusives* (*nasales* comprises) dominent nettement et les *fricatives* ne sont représentées que par /s/ (et /h/, classé fricatif par Maddieson comme par l'API, mais dont le statut est pour le moins spécifique, cf. le nœud laryngal dans la géométrie des traits). Les fricatives sont donc bien les plus rares dans le babillage et c'est aussi ce que l'on vient d'observer dans la base des langues du monde. À partir de cette tendance – parmi d'autres –, Locke et Pearson (1992, p. 96) n'hésitent pas à tirer, à juste titre, la conclusion de portée générale suivante :

> « Infants heavily favor stop consonants over fricatives, and there are languages that have stops and no fricatives but no languages that exemplify the reverse pattern. [Such] "phonologically universal" patterns, which cut across languages and speakers are, in effect, the phonetic properties of *Homo sapiens*. »

Table 2. Phonèmes consonantiques les plus fréquents ($UPSID_{451}$ N = 10.146 consonnes). /t d n l/ regroupent toutes les coronales.

C	%	C	%
t	97.5	g	56.2
m	94.4	ŋ	52.7
n	90.4	ʔ	48.0
k	89.5	tʃ	41.8
j	84.0	ʃ	41.6
p	83.3	f	40.0
w	76.8	dz	34.9
s	73.5	ɲ	31.3
d	64.7	ts	29.3
b	63.8	k^h	22.9
h	62.0	p^h	22.4
l	56.9	v r	21.1

4. CONCLUSIONS ET PERSPECTIVES

Les typologies phonologiques et les tendances universelles mises en évidence sont orthogonales aux familles linguistiques qui ont servi pour l'échantillonnage de la base. Tout au plus peut-on remarquer que certains traits (arrondissement, longueur, nasalité...) sont associés à certaines familles ou plutôt à des aires géographiques. Ceci ne fait que confirmer que le classement *typologique* et le classement

génétique sont loin d'être identiques, et que la typologie a des tendances *géographiques*. Cet état des choses a fait l'objet de nombreuses remarques depuis Jakobson (1936) ; cf. Ruhlen (1978) pour les nasales. Il est vraisemblable que la statistique *au cours du temps*, dans une zone déterminée de l'espace linguistique, reflète les mêmes contraintes que la distribution générale actuelle, à ceci près, qu'en ce qui concerne certains traits, les fluctuations des systèmes des zones voisines se produiraient plutôt en phase.

Une des inférences intéressantes de ce manque de relations est que les tendances générales des systèmes pourraient dépendre de la *matérialité* des structures sonores du langage ; et il semble qu'on puisse en trouver une confirmation dans l'*ontogenèse*, appuyant ainsi la thèse qu'existeraient des « propriétés phonétiques d'*Homo sapiens* ».

Dans un second volet de cette étude, nous tenterons une corroboration de ce type d'hypothèse en montrant, par la théorie de la dispersion focalisation (TDF), comment certaines des grandes tendances universelles peuvent être paramétrisées (Boë & al., 1996).

4. RÉFÉRENCES

Bernatova N., Zackova P. (1992) *Typologie des occlusives et des constrictives à partir de la base de données* UPSID. DEA Univ. Stendhal, Grenoble.

Berrah A.R. (1994) L'émergence des structures sonores : les syllabes consonnes/voyelles. DEA Sciences Cognitives, Grenoble.

Boë L.J., Schwartz J.L., Vallée N., Abry C. & Berrah R. (1996) La matérialité des structures sonores du langage 2. De la prédiction à l'ontogenèse. 21e JEP, dans ces mêmes Actes.

Crothers J. (1978) Typology and Universals of Vowel Systems. In *Universals of Human Language*. J.H. Greenberg (Ed.), 93-152, Stanford Univ.Press.

Greenberg J.H., Ferguson C.A., Moravcsik E.A. (Ed.) (1978) *Universals of Human Languages: Method and Theory, Phonology, Word Structure, Syntax.* Stanford Univ. Press.

Hockett C.F. (1955) *A Manual of Phonology*. Waverly Press, Baltimore, 246p. *Publications in Anthropology and Linguistics*, Indiana University, Bloomington.

Jakobson R. (1936) Sur les affinités phonologiques entre les langues. *4e Congrès International de Linguistes*, 48-58.

Jakobson R. (1941) *Kindersprache, Aphasie, und allgemeine* Lautgestze. Uppsala Universitets Årsskrift 1942, 1-83. Republié in *Selected writings I.* 1962, Mouton, The Hague, 328-401.

Jomaa M., Abry C. (1992) *La base de systèmes vocaliques* RHÔNSON. Rapport PPSH 78B.

Ladefoged P. (1995) The Sounds of Disappearing Languages. *The newsletter of The Acoustical Society of America*, 5, 1, 1-6.

Liljencrants J., Linblom B. (1972) Numerical Simulation of Vowel Quality Systems: the Role of Perceptual Contrast. *Language* 48, 839-862.

Lindblom B., Maddieson I. (1989) Phonetic Universals in Consonant Systems. In *Language, Speech and Mind. Studies in Honour of Victoria A. Fromkin.* L.M. Hyman & C.N. Li. (Eds.) Routledge, London and New-York.

Locke J.L. (1983) *Phonological acquisition and change.* Academic Press, New-York.

Locke J.L., Pearson D.M. (1992) Vocal Learning and the Emergence of Phonological Capacity. A Neurobiological Approach. In *Phonological Development. Models, Research, Implications.* C. Ferguson, L. Menn & C. Stoel-Gammon. (Eds.) Timomium, Maryland, 91-129.

Maddieson I. (1984). *Patterns of sounds.* Cambridge University Press, Cambrige (2e édition 1986).

Maddieson I. (1991) Testing the Universality of Phonological Generalizations with a Phonetically Specified Segment Database: Results and Limitations. *Phonetica* 48, 193-206.

Maddieson I. & Precoda K. (1989) Updating UPSID. *UCLA WPP* 74, 104-111.

Ruhlen M. (1978) Nasal vowels. In *Universal of Human Language.* J. Greenberg & al. (Eds.), vol. 1, 203-241.

Ruhlen M. (1987) *A Guide of the World's Languages : Classification.* Arnold E. (Ed.), London.

Sedlak P. (1969). Typological Considerations of Vowel Quality Systems. *Working Papers on Language Universals* 1, Stanford University, 1-40.

Schwartz J.L., Boë L.J., Vallée N., Abry C. (1996) The Dispersion-Focalization Theory of Vowel Systems. *J. of Phonetics*, soumis.

Stevens K.N. (1972). The Quantal Nature of Speech: Evidence from articulatory-acoustic data. In *Human Communication: A unified view,* E.E.Davis Jr. & P.B. Denes (Eds.) 51-66. New-York: Mc Graw-Hill.

Troubetzkoy N.S. (1939). Grundzüge der Phonologie. *Travaux du Cercle Linguistique de Prague* 7, 272p. Trad. française par J. Cantineau *Principes de phonologie.* Klincksieck, 1949, 1970, Paris.

Vallée N. (1989) *Typologie des systèmes vocaliques.* T.E.R. de Maîtrise, Université Stendhal, Grenoble.

Vallée N. (1994) *Systèmes vocaliques : de la typologie aux prédictions.* Thèse de Doctorat en Sciences du Langage, Université Stendhal, Grenoble.

Vallée N., Boë L.J., Schwartz J.L. (1990) Systèmes vocaliques : typologie et tendances universelles. *18e Journées d'Étude sur la Parole,* 32-36.

REMERCIEMENTS

Un grand merci à Ian Maddieson qui a mis à notre disposition $UPSID_{451}$ et qui a consacré beaucoup de son temps à répondre à nos questions ; à Jean-Marie Hombert qui nous a ouvert sa bibliothèque « californienne » ; à Denis Creissels pour ses nombreuses et précieuses informations ; à Carol Stoel-Gammon pour nous avoir fait bénéficier, pendant son séjour à l'ICP, de ses grandes connaissances sur l'acquisition du langage.

La matérialité des structures sonores du langage
2. De la prédiction à l'ontogenèse

Louis-Jean Boë, Jean-Luc Schwartz, Ahmed Berrah, Nathalie Vallée, Christian Abry

ICP UPRES A 5009 INPG/Université Stendhal, BP 25 38040 Grenoble Cedex 9
Tél : 76 82 43 38 – Fax : 76 82 43 35 – e-mail (boe, schwartz, berrah, vallée, abry)@icp.grenet.fr

ABSTRACT

The main tendencies of sound structures of the world languages (Vallée & al., 1996) seem to us a good framework to test the influences of constrains of production and perception on phonological systems. We present here results of prediction of vowel and syllable systems by the Theory of Dispersion Focalization of ICP. At last perpectives are proposed to discuss results in the light of ontogeny.

1. LE CADRE ÉPISTÉMOLOGIQUE

Nous avons choisi un cadre épistémologique (figure 1) dans lequel la faiblesse intrinsèque de la seule démarche inductive – pouvant entraîner l'erreur d'affirmer le conséquent d'observations partielles – est compensée par des principes et des données de production et de perception externes au langage du point de vue formel, c'est-à-dire indépendantes des contraintes purement cognitives du langage. C'est le cadre défini par Lindblom & al. (1996) pour avancer l'hypothèse selon laquelle c'est la substance qui structure la forme et non l'inverse. Nous introduirons comme données externes les capacités psychophysiques de contrôle des dimensions vocaliques principales, par la langue et la mandibule ; ainsi que l'intégration perceptive large bande (ou F'_2). Les données typologiques sur lesquelles sont basées nos prédictions ont été obtenues à partir des statistiques sur les inventaires phonémiques et syllabiques.

2. MODÉLISATION DE LA PRÉDICTION DES STRUCTURES SONORES

2.1. Les voyelles

Depuis Liljencrants & Lindblom (1972) différentes hypothèses ont été avancées pour expliquer la structure sonore des systèmes vocaliques à partir de contraintes articulatoires et auditives. Les meilleurs résultats ont été obtenus avec la théorie de la dispersion adaptative (TD) de Lindblom (1986). Mais les prédictions ainsi réalisées présentent un trop grand nombre de voyelles hautes non périphériques /i ʉ ɯ/ entre /i/ et /u/.

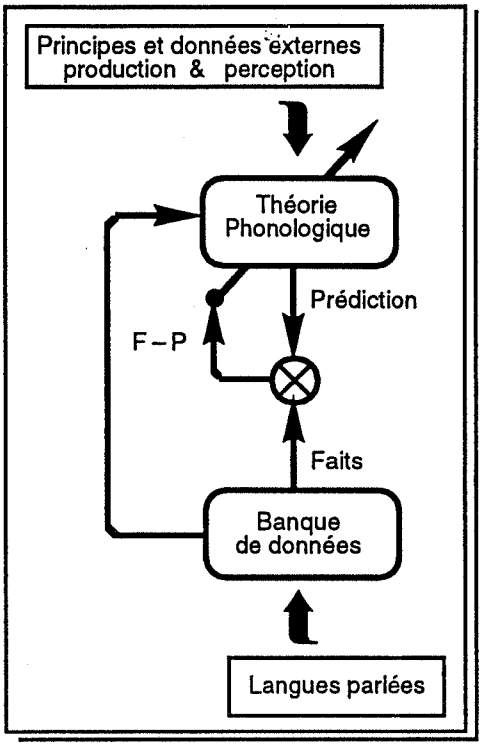


Figure 1 : Le cadre épistémologique.

La théorie ICP de la dispersion-focalisation (TDF) (Schwartz et al., 1989 ; Abry et al., 1994) permet de résoudre ce problème en mettant en compétition, avec le principe de *distance* entre voyelles, base de la TD, un principe gestaltiste de *prégnance* intravoyelle : la *focalisation*. Dans l'espace vocalique des formants, la proximité de deux résonances – *convergence articulatoirement produite autour d'un changement d'affiliation entre cavités* (Abry & al., 1989 ; Badin & al., 1990) –, a pour effet de renforcer l'énergie dans une zone du spectre, donnant ainsi à la voyelle une qualité « focale » que l'on peut comparer à celle des couleurs qualifiées ainsi (Brown & Lenneberg, 1954), sur lesquelles s'accordent les perceptions des sujets de différentes langues (Rosch-Heider, 1972).

La théorie de la dispersion-focalisation TDF fait donc l'hypothèse que, pour un nombre de voyelles donné n, le système phonologique préféré (c'est-à-dire le plus fréquent dans une base représentative des langues du monde) est obtenu par minimisation d'un coût d'énergie global :

$$E_{DF} = E_D + \alpha E_F$$

avec la composante de dispersion structurelle :

$$E_D = \sum\sum (1/d_{ij})^2 \quad (i=1..n-1 \quad j=i+1..n)$$

et la composante locale de focalisation E_F pondérée par le paramètre α :

$$E_F = (E_{12} + E_{23} + E_{34})$$

avec

$$E_{f\ f+1} = -\sum\sum 1/(F^i_{f+1} - F^i_f)^2 \quad i = 1..\ n\ f = 1..3$$

Les voyelles V^i sont décrites par 4 formants (F_4 étant fixé à 3650 Hz) exprimés en barks selon la formule proposée par Schroeder & al. (1979) : bark = 7 ArgSh (Hz / 650).

Pour calculer la distance perceptive d_{ij} entre deux voyelles V^i et V^j, on utilise une distance euclidienne sur F_1 et F'_2. Le second formant perceptif est évalué à partir de F_2, F_3 et F_4 sur la base d'un modèle qui a été introduit par Fant et développé par la suite (Schwartz, 1987). Nous avons introduit un poids λ de F'_2 par rapport à F_1 :

$$d_{ij} = [(F_1^j - F_1^i)^2 + \lambda^2 (F'^j_2 - F'^i_2)^2]^{1/2}$$

pour tenir compte de la dilatation du triangle vocalique dans le sens de F_1 par rapport à F_2 (Lindblom, 1986), qui pourrait provenir soit de mécanismes psycho-acoustiques de masquage des aigus par les graves, soit de différences de capacités psychomotrices du contrôle vertical de la mandibule par rapport au déplacement avant-arrière de la langue (Boë & al., 1994).

La TDF se différencie donc de la TD par l'introduction d'un deuxième coût qui diminue l'énergie des systèmes pour lesquels les voyelles ont des formants F_1 et F_2, F_2 et F_3 ou F_3 et F_4 rapprochés – voyelles « focales » – ce qui les rend plus stables.

La TDF est ainsi paramétrisée par λ et α qui renvoient à des propriétés externes, au sens où nous les avons définies en introduction. Pour tester leurs valeurs nous avons adopté une méthode utilisée notamment en chimie : la détermination de l'espace des phases (Schwartz & al., 1995). Les différentes zones des espaces de phase ont été explorées extensivement, en particulier, grâce à une étude utilisant une méthode d'optimisation par algorithme génétique (Fargetton, 1993). Il s'est agi de trouver une zone λ et α permettant de prédire les systèmes préférentiels de 3 à 9 voyelles. Les résultats montrent que cette zone existe, et qu'elle est délimitée par :

$$0.2 \leq \lambda \leq 0.3 \text{ et } 0 \leq \alpha \leq 0.4$$

avec, pour assurer la stabilité des systèmes comportant la voyelle /y/, une contrainte supplémentaire sur α : $0.3 \leq \alpha \leq 0.4$. On retrouve bien une valeur de λ voisine de celle mise en évidence par les tests de DME (*Direct Magnitude Estimation*) de Lindblom & Lubker (1985). Si la focalisation semble une propriété du système auditif liée à des propriétés générales d'intégration (Schwartz & Escudier, 1989), il reste à justifier – psycho-physiquement, informationnellement et/ou cognitivement – la valeur de la pondération α, c'est-à-dire le poids de la focalisation par rapport à la dispersion.

2.2. Les syllabes

Sur les syllabes de lexiques à 2000 entrées de 31 langues, ce sont les syllabes CV qui sont les plus nombreuses (Janson, 1986 ; Maddieson, 1992). Remarquons, au passage, que ce sont aussi les syllabes canoniques du babillage de l'enfant. Dans les langues du monde, les occlusives /p b t d k g/ sont les consonnes les plus fréquentes, les occlusives pharyngales /ʔ/ n'étant que très peu attestées (0.7% dans $UPSID_{450}$). Comme /i a u/ sont les voyelles de loin les plus fréquentes, suivies de /'e' 'o'/ on peut avancer, sans trop prendre de risques, que /ba da ga bi di gi bu du gu/ sont les syllabes les plus fréquentes suivies de /b'e' d'e' g'e' b'o' d'o' g'o'/. Nous avons limité notre simulation aux 20 syllabes CV combinaisons de C = /b d g ʔ/ et V = /i 'e' a 'o' u/. Les prototypes des cibles articulatoires consonnes et voyelles coarticulées ont été obtenues par expertise (Vallée, 1994 ; Vallée & al., 1995) et par inversion (Bailly & al., 1995) à l'aide du modèle de Maeda (1989). L'articulatoire des prototypes a été confrontée aux données radiographiques qui ont servi à l'élaboration du modèle (Bothorel & al., 1986) et l'acoustique aux locus de Sussman & al. (1993). Chaque syllabe CV est donc identifiée par ses caractéristiques articulatoires (paramètres du modèle articulatoire choisi) et acoustiques (F_1 et F'_2). Le modèle prédictif (Berrah, 1994 ; Berrah & al., 1995) des syllabes repose sur la TD et il introduit la notion d'efficacité globale (*Global Efficiency*). Un système de n syllabes donné minimise :

$$\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} \frac{1}{d_{s_i s_j}^{\ 2}} + \beta \sum_{i=1}^{n} \left(\frac{1}{\text{Glob Eff}_{s_i}} \right)$$

La distance intersyllabique perceptive d_{S1S2} ($S_1 = C_1V_1$, $S_2 = C_2V_2$) est décrite par :

$$\sqrt{(F_{1_{c1}} - F_{1_{c2}})^2 + \lambda^2 * (F'_{2_{c1}} - F'_{2_{c2}})^2 + (F_{1_{v1}} - F_{1_{v2}})^2 + \lambda^2 * (F'_{2_{v1}} - F'_{2_{v2}})^2}$$

avec la pondération λ entre F_1 et F'_2. L'efficacité globale d'une syllabe est définie par le rapport de son efficacité acoustique (*salience*) à son coût articulatoire :

$$\text{Glob Eff}_{cv} = \frac{\text{Acoust Eff}_{cv}}{\text{Art Cost}_{cv}}$$

L'efficacité acoustique intrasyllabique Eff_{cv} est définie par l'excursion de la transition CV :

$$\sqrt{(F_{1_c} - F_{1_v})^2 + \lambda^2 * (F'_{2_c} - F'_{2_v})^2}$$

Le coût articulatoire d'une syllabe CV produite par le jeu des m paramètres articulatoires P_k (m =7 dans le cas du modèle Maeda) est donné par :

$$\text{Art Cost}_{CV} = \sqrt{\sum_{k=1}^{m} \omega_p^2 * (P_{kc} - P_{kv})^2}$$

Le paramètre β pondère le coût intrasystémique par rapport au coût intersystémique. Nous avons ajusté le poids des différents paramètres ω_p en tenant compte des données sur la langue et la mandibule et des statistiques sur les syllabes. Par exemple Maddieson & Precoda (1992) ont montré que, sur un vaste échantillon de langues, les syllabes /di/ et /du/ ont les mêmes occurrences – ce que nous avons traduit par même efficacité globale –, alors que leurs coûts articulatoires et leurs efficacités acoustiques sont très différentes. Après ajustement de tous les paramètres, émergent par ordre d'efficacité globale : /b'o' bu d'e' g'e' b'e' ba ga g'o' da gi bi gu du di ʔa do ʔ'o' ʔi ʔ'e' ʔu/. Mais si on choisit un système à 9 syllabes parmi ces 20 syllabes c'est bien le système /bu ba ga da gi bi gu du di/ qui apparaît alors.

3. CONCLUSIONS ET PERSPECTIVES

La prédiction des tendances universelles des systèmes vocaliques et syllabiques CV (même si cette dernière est bien moins avancée) confirme la viabilité du projet d'une *linguistique orientée substance*. Resituer cette approche dans le cadre de l'ontogenèse permet de l'enrichir d'une nouvelle dimension et de la prolonger grâce aux données d'un domaine en plein développement (cf. par. exemple Locke, 1983 ; Ferguson & al., 1995). Les langues constitueraient leur matériau sonore phonologique universel à partir des potentialités articulatoires CVCV d'un générateur de babillage (figure 2). Le lexique de l'enfant pourrait être considéré comme le produit d'un filtrage et d'un ajustement (*tuning*) de ses CVCV par le milieu maternel avec une interaction auditive et visuelle. Le découpage en C et V opéré par l'analyse phonologique ne serait enfin qu'une opération formelle sur des unités initiales syllabiques qui en délivrerait ainsi un sous-produit. Reste à montrer qu'une analyse phonologique du lexique en syllabes est plus riche en prédiction qu'une analyse opérée à partir des constituants vocaliques et consonantiques.

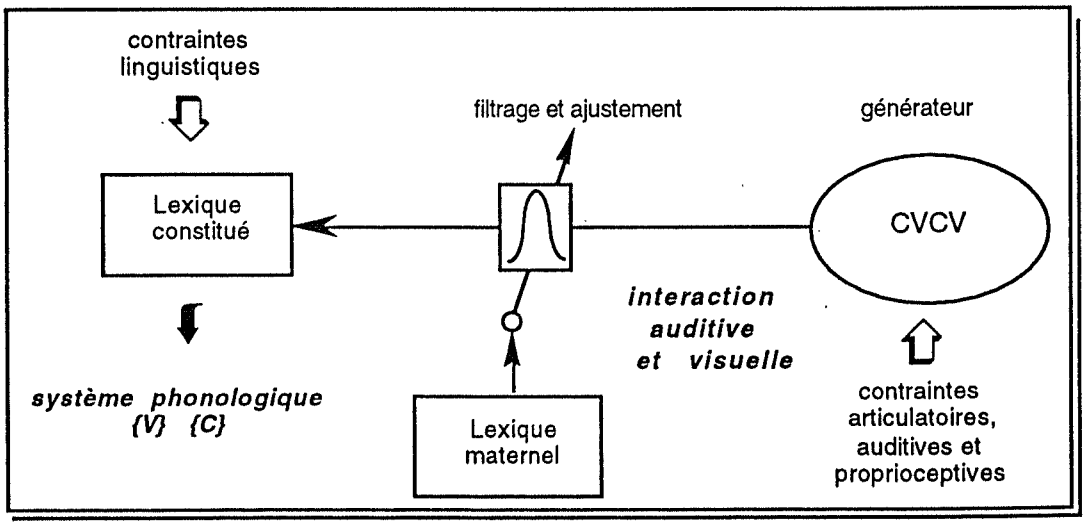


Figure 2 : Le lexique constitué de l'enfant considéré comme le résultat du filtrage-ajustement de son babillage CVCV par l'environnement maternel.

4. RÉFÉRENCES

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., Boë L.J., Vallée N., Schwartz J.L. (1994) La théorie ICP de la dispersion-focalisation à l'épreuve des systèmes vocaliques des langues du monde et du terrain francoprovençal. *Colloque Terrain et Théorie en Linguistique*, Paris.

Bailly G., Boë L.J., Vallée N. (1995) Articulatori-Acoustic Vowel Prototypes for Speech Production. *Eurospeech 95* 3, 1913-1916.

Badin P., Boë L.J., Perrier P., Abry C. (1990) Vocalic Nomograms : Acoustic and Articulatory Considerations upon Formant Convergences. *J. Acoust. Am.* 87, 1290-1300.

Berrah A.R. (1994) *L'émergence des structures sonores : les syllabes consonnes/voyelles.* DEA Sciences Cognitives. INP Grenoble.

Berrah A.R., Boë L.J., Schwartz J.L. (1995) Emergent Syllable using Articulatory Acoustic Principles. *XIIIth Int. Congr. of Phonetic Sciences*, 1, 396-399.

Boë L.J., Schwartz J.L., Vallée N. (1994) The Prediction of Vowel Systems : Perceptual Contrast and Stability. In *Fundamentals of Speech Synthesis and Speech Recognition.* E. Keller (Ed.), John Wiley, London, England.

Bothorel A., Simon P., Wioland F., Zerling J.P. (1986) *Cinéradiographies des voyelles et des consonnes du français.* Institut de Phonétique, Strasbourg.

Brown R.W., Lenneberg E.H. (1954) A Study in Language and Cognition. *J. Abnormal and Social Psychology* 49, 454-462.

Fargetton L. (1993) *Prédiction des systèmes vocaliques par algorithmes génétiques.* Stage DUT, IUT Informatique Grenoble 2, ICP.

Ferguson, C. Menn L. & C. Stoel-Gammon. (Eds.) (1995) *Phonological Development. Models, Research, Implications.* Timomium, Maryland,

Janson T. (1986) Cross-linguistic Trends in the Frequency of CV Sequences. *Phonology Yearbook* 3, 179-195.

Liljencrants J., Lindblom B. (1972) Numerical Simulation of Vowel Quality Systems : The Role of Perceptual Contrast. *Language* 48, 839-862.

Lindblom B. (1986) Phonetic Universals in Vowel Systems. In *Experimental Phonology.* J.J. Ohala (Ed.), Academic Press, Orlando, Florida.

Lindblom B., Lubker J. (1985). The Speech Homonculus : A Problem of Phonetic Linguistics. In *Phonetic Linguistics.* V.A. Fromkin (Ed.), Academic Press, Orlando, Florida, 169-192.

Lindblom B., MacNeilage P., Studdert-Kennedy M. (1996) *Evolution of Spoken Language*, chap. 7-9.(à paraître)

Locke J.L. (1983) *Phonological acquisition and change.* Academic Press, New-York.

Maddieson I. (1984) *Patterns of sounds.* Cambridge University Press (2e éd. 1986).

Maddieson I. (1992) The Structure of Segment Sequences. *UCLA Working P. in Phonetics* 83, 1-7.

Maddieson I., Precoda K. (1992) Phonetic models and syllable structure, *Phonology* 9, 45-60.

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.

Rosch-Heider E. (1972) Universals in Color Naming in Memory. *J. Experimental Psychology* 93, 10-20.

Schroeder M.R., Atal B.S., Hall J.L. (1979). Objective Measure of certain Speech Signal Degradations based on Masking Properties of Human Auditory Perception. In *Frontiers of Speech Communication Research*, B. Lindblom, S. Ohman, (Eds.) 217-229. London: Academic Press.

Schwartz J.L. (1987) *Représentations auditives des spectres vocaliques.* Thèse de Docteur es Sciences Physiques, INP Grenoble.

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* 1, 63-66.

Schwartz J.L., Boë L.J., Vallée N. (1995) Testing the Dispersion-Focalization Theory: Phase Spaces for Vowel Systems. *XIIIth Int. Congr. of Phonetic Sciences*, 1, 412-415.

Schwartz J.L., Boë L.J., Vallée N., Abry C. (1996) The Dispersion-Focalization Theory of Vowel Systems. *J. of Phonetics*, soumis.

Schwartz J.L., Escudier P. (1989) A Strong Evidence for the Existence of a Large Scale Integrated Spectral Representation in Vowel Perception. *Speech Communication*, 8, 235-259.

Sussman H.M., Hoemeke K.A., Farhan S.A. (1993) A Cross-linguistic Investigation of Locus Equations as a Phonetic Descriptor for Place articulation, *J. Acoust. Soc. Am.*, vol. 94, 1256-1268.

Vallée N. (1994) *Systèmes vocaliques : de la typologie aux prédictions.* Thèse de Doctorat en Sciences du Langage, Université Stendhal, Grenoble.

Vallée N., Abry C., Boë L.J., Schwartz J.-L., Berrah A. (1996) La matérialité des structures sonores du langage 1. Taxinomies phonologiques et tendances universelles. *Dans ces mêmes Actes.*

Vallée N., Boë L.J., Payan Y. (1995) Vowel prototypes for UPSID'S phonemes. *XIIIth Int. Congr. of Phonetic Sciences* 1, 424-427.

DETTES ET RECONNAISSANCES

Un grand merci à Bjorn Lindblom pour nous avoir communiqué certains de ses travaux *in Press* et de précieuses références bibliographiques ; Pierre Bessière pour son encadrement dans l'exploitation des algorithmes génétiques ; Shinji Maeda et Ken Stevens pour leurs judicieux conseils et critiques constructives.

Une partie de cette recherche a été élaborée, ou présentée dans le cadre du projet européen ESPRIT BRA *Speech Maps* (*Primes* Christian Abry et Pierre Badin) ; elle a été soumise à discussion au cours des séminaires croisés (1994/95) entre le laboratoire Dynamique du Langage à Lyon et l'ICP à Grenoble.

LE TRAITEMENT DES EXPRESSIONS IDIOMATIQUES AMBIGUËS

Maryline NGUYEN, Pierre MARQUER et Juan SEGUI

Laboratoire de Psychologie Expérimentale, Université René Descartes (Paris V) - U.R.A. 316 du C.N.R.S.

28, rue Serpente, 75006 Paris.

Tél.: 40 51 98 65 - Fax: 40 51 70 85 - e-mail: crea3@idf.ext.jussieu.fr

ABSTRACT

Three cross-modal priming experiments were conducted to investigate the activation of idioms' literal meanings and its relation to idiomatic access. Idioms were embedded in neutral sentence contexts, and three target types were used. « Idiomatic » targets were related to the figurative meanings of the presented idioms; « literal » targets, to their overall literal interpretations, and « last word targets », to the literal meaning of each idiom's last word. When targets were presented 300 ms before idiom offset (Experiment 1), only the global literal interpretation of idioms was activated. Immediately after idiom offset (Experiment 2), last word and idiomatic activation emerged, but literal activation was not significant. When the test-point was shifted to 300 ms after idiom offset (Experiment 3), literal activation reappeared but, although idioms' last words were activated, there was no evidence of the idiomatic meaning's being available. These data are discussed within the framework of a recent model of idiom processing, in which idiomatic and literal interpretations are derived in parallel, without activation of the former stopping the ongoing literal analysis. The present results furthermore suggest that although the literal interpretation cannot be suspended, it could momentarily cease to be available.

1. INTRODUCTION

Les expressions idiomatiques (ex: « Casser sa pipe ») sont traditionnellement définies comme des énoncés complexes dont le sens conventionnel (« mourir ») n'est pas le résultat direct d'une analyse syntaxico-sémantique des mots qui les composent (une telle analyse conduirait ici à comprendre « Casser sa pipe » comme « briser un instrument servant à fumer »).

Une analyse « littérale » semblant donc inutile, les premiers modèles psycholinguistiques rendant compte de la compréhension des expressions idiomatiques postulaient que le calcul d'une interprétation littérale, qu'il soit réalisé (Swinney & Cutler, 1979) ou pas (Gibbs, 1980), n'intervient en aucune façon dans le processus d'accès au sens idiomatique, conçu comme la simple récupération d'une signification stockée « en bloc » dans le lexique mental.

Cependant, des résultats plus récents suggèrent que le traitement des expressions idiomatiques présente en réalité de nombreuses similarités avec celui des énoncés littéraux: non seulement un accès au sens des composants individuels de l'expression (Gibbs & Nayak, 1989), mais aussi une analyse syntaxique complète de celle-ci (voir Peterson, Burgess, Dell, & Eberhard, 1989, cités par Peterson & Burgess, 1993), seraient réalisés de façon automatique lors de la compréhension de ces expressions.

Cela signifie-t-il qu'une interprétation littérale globale est également dérivée?

Deux recherches récentes ont donné à cette question une réponse affirmative, toutefois assortie de certaines conditions. Pour Peterson *et al.* (1989), bien que le traitement syntaxique de l'expression soit obligatoire, son interprétation littérale n'est activée que si le contexte dans lequel elle est présentée favorise un traitement non idiomatique. Titone & Connine (1994) insistent quant à elles, non pas sur le rôle des informations contextuelles, mais sur celui de certaines caractéristiques intrinsèques des expressions: moins une expression est prédictible et plus son interprétation littérale est plausible, plus grande est la probabilité que cette dernière soit activée.

La possibilité que l'interprétation littérale des expressions idiomatiques soit activée au cours de leur traitement appelle une seconde question: quelles sont les relations existant entre l'activation des deux interprétations possibles d'une expression donnée? Plus spécifiquement, l'activation d'une des deux interprétations possibles peut-elle favoriser, ou, au contraire, gêner, l'activation de l'autre interprétation?

De précédents travaux ont montré que l'accès au sens idiomatique pouvait suspendre l'analyse littérale d'une expression (Swinney, 1981; Peterson & Burgess, 1993) et qu'inversement, la présence d'une interprétation littérale était susceptible de réduire la dis-

ponibilité de son sens idiomatique (Titone & Connine, 1994).

Cependant, certains problèmes méthodologiques limitent la portée de ces résultats. En effet, soit les méthodes employées ne reflètent qu'imparfaitement les phases précoces des phénomènes étudiés (Swinney, 1981), soit, pour la plupart des recherches dites « en temps réel » (Cacciari & Tabossi, 1988; Colombo, 1993; Titone & Connine, 1994), l'activation du sens littéral global des expressions est confondue, sinon conceptuellement, du moins expérimentalement, avec celle de leur dernier mot, puisque les cibles dites « littérales » n'ont, dans ces travaux, de lien sémantique qu'avec le dernier mot de l'expression.

La présente recherche se propose donc, tout en utilisant un paradigme d'amorçage intermodal classique avec tâche de décision lexicale (voir par exemple Cacciari & Tabossi, 1988), d'étudier l'éventualité d'une activation de l'interprétation littérale des expressions idiomatiques de façon directe, en introduisant des cibles liées au sens littéral global des expressions présentées (ex: « Passer la corde au cou »-PENDRE), en plus des cibles classiquement utilisées, liées au sens littéral de leur dernier mot (ex: NUQUE).

Le décours temporel des relations possibles entre les deux interprétations des expressions idiomatiques sera par ailleurs étudié en comparant l'effet d'amorçage obtenu pour les cibles liées au sens littéral global des expressions à celui observé pour des cibles liées à leur sens figuré global (ex: « Passer la corde au cou »-ÉPOUSER), ceci à trois moments différents: immédiatement à la fin de l'expression, 300 ms avant sa fin, et 300 ms après.

2. EXPÉRIENCES

2.1. Sujets

Chacune des trois expériences décrites a été passée par 45 étudiants en Sciences Humaines de l'Université René Descartes (Paris V), soit un total de 135 sujets. Tous étaient de langue maternelle française, avaient une vision normale ou corrigée, et ne présentaient aucun déficit auditif connu. Aucun d'entre eux n'était informé de la présence d'expressions idiomatiques dans le matériel expérimental avant la fin de la passation.

2.2. Matériel

15 expressions ambiguës (possédant une interprétation littérale plausible) composées d'un verbe et d'un groupe nominal et/ou d'un groupe prépositionnel, ont été choisies dans un dictionnaire d'expressions et locutions françaises (Rey & Chantreau, 1991).

A chacune d'entre elles sont associés trois mots-cibles: une cible « idiomatique », sémantiquement associée au sens figuré global de l'expression correspondante (ex: « Passer la corde au cou »-ÉPOUSER), une cible « littérale », liée à une interprétation littérale globale de cette expression (PENDRE), et une cible « dernier mot » liée au sens littéral de son composant final (NUQUE).

Le choix des cibles littérales et idiomatiques a été effectué à partir des productions de 20 sujets auxquels il a été demandé, au cours d'une préexpérience, de donner une définition du sens figuré, d'une part, littéral, d'autre part, de chacune des expressions retenues. Le choix des cibles liées au dernier mot des expressions a été réalisé sur la base d'associations sémantiques. L'existence d'un lien sémantique entre chaque expression et les trois types de cibles associés a été vérifiée auprès de 5 juges indépendants. La fréquence d'usage des cibles n'a pas été prise en compte en tant que telle, le plan expérimental (voir section 2.3) étant construit de façon à ce que chaque cible soit son propre contrôle.

Les expressions sont présentées à l'infinitif et dans leur forme canonique, à la fin d'une phrase-contexte (comprenant de 9 à 16 mots) n'induisant ni une interprétation littérale, ni une interprétation idiomatique (ex: « Il espérait ainsi reculer le moment où on allait lui passer la corde au cou. »). Les phrases-contextes expérimentales sont mélangées à 105 phrases dites « de remplissage », qui leur sont appariées en longueur et structure.

Une présentation auditive des phrases-amorces a été choisie de préférence à une présentation visuelle, non seulement pour que les sujets ne puissent choisir d'ignorer les phrases-amorces (ce qui aurait été possible dans la modalité visuelle), mais également pour permettre la présentation de cibles avant la fin de l'amorce.

Les phrases ont été enregistrées par une locutrice de langue maternelle française sur un magnétophone digital (D. A. T.). La prononciation a été maintenue aussi neutre que possible, de façon à minimiser l'éventuelle influence de facteurs prosodiques sur le traitement des expressions présentées. La durée moyenne des phrases-amorces est de 3600 ms (de 2400 à 4200 ms). Au sein de ces dernières, les expressions idiomatiques elles-mêmes durent en moyenne 870 ms (de 650 à 1100 ms).

2.3 Listes expérimentales

Afin que tous les sujets soient confrontés à tous les types de liaison amorce-cible, mais que chaque item (amorce ou cible) ne soit présenté

qu'une seule fois au cours d'une session expérimentale, trois listes ont été construites.

Dans chaque liste, seul un tiers des cibles expérimentales est effectivement précédé de la phrase contenant l'expression correspondante, les deux tiers restants étant présentés après une phrase de remplissage avec laquelle ils ne possèdent aucun lien sémantique. Afin de maximiser l'équivalence *a priori* des listes, les cibles ont été réparties entre celles-ci sur la base des temps moyens de décision lexicale recueillis pour chaque cible (présentée sans l'amorce correspondante) au cours d'une préexpérience.

Au sein de chacune des listes, 30 des phrases de remplissage sont suivies d'une cible expérimentale. 15 autres sont associées à un mot-cible « de remplissage » sémantiquement relié (ex: « Il avait essayé par tous les moyens de perdre du poids »-RÉGIME). Les 60 dernières sont suivies d'un non-mot (ex: MARINDE), afin de permettre la réalisation de la tâche de décision lexicale.

2.3. Procédure expérimentale

Une procédure similaire a été utilisée lors des trois expériences. Au cours d'une session expérimentale individuelle d'environ 20 mn, la série de phrases-amorces est présentée auditivement à partir du disque dur d'un ordinateur. Après chaque phrase, une cible visuelle apparaît sur l'écran de l'ordinateur. La tâche des sujets est alors de décider le plus rapidement et le plus exactement possible si cette cible est un mot de la langue française ou non (décision lexicale), en appuyant sur un des deux boutons prévus à cet effet. On enregistre la latence (en ms) et la nature de chaque réponse.

Afin de s'assurer que les sujets ont, conformément aux instructions qui leur sont données, prêté une attention suffisante aux phrases-amorces, on leur demander d'effectuer, à l'issue de la passation, une épreuve de reconnaissance sur un échantillon de 12 phrases extraites du matériel, mélangées à 12 autres phrases « nouvelles ». Seuls les protocoles des sujets capables d'identifier correctement un minimum de 50% des phrases ayant fait partie de l'expérience sont conservés pour l'analyse finale.

L'unique différence entre les trois expériences réside dans le moment d'apparition de la cible visuelle par rapport à la phrase-amorce. Dans la première expérience, la cible est présentée 300 ms avant la fin de la phrase; dans la deuxième, elle apparaît immédiatement après celle-ci; dans la troisième, enfin, la cible est présentée 300 ms après la fin de l'amorce auditive.

Table 1 - Effets d'amorçage en ms en fonction du type de cible et du délai entre la phrase-amorce et la cible correspondante. ** signale un effet significatif (à un seuil d'au moins 0,05) par sujets et par items; *, un effet significatif par sujets uniquement; NS, un effet non significatif.

	Délai amorce-cible		
Type de cible	-300 ms (exp. 1)	0 (exp. 2)	+300 ms (exp. 3)
Idiomatiques	+5 NS	+19 *	+2 NS
Littérales	+33 **	+19 NS	+24 **
« Dernier mot »	+14 NS	+21 **	+30 **

2.4. Résultats

Les expériences présentées ayant été conçues pour étudier des effets d'amorçage (c'est-à-dire des différences de temps de réponse entre les conditions où il existe un lien amorce-cible et les conditions où ce lien est absent), seuls ces effets pour les trois types de cibles seront exposés (Table 1) et discutés.

Les résultats obtenus montrent que 300 ms avant et après la fin des expressions présentées, les cibles « littérales » sont traitées plus rapidement (de 33 et 24 ms respectivement) si elles sont précédées de l'expression correspondante que lorsque ce n'est pas le cas. Ceci suggère qu'une interprétation littérale globale est effectivement calculée lors du traitement des expressions idiomatiques ambiguës et que cette interprétation est disponible, du moins lorsque les expressions sont présentées, comme ici, dans un contexte ne contraignant pas les sujets à les interpréter de façon figurée. Il semble de plus que cette interprétation littérale soit activée de façon précoce, et se maintienne relativement tardivement.

Notons que, au moins dans la première expérience, des résultats différents ont été obtenus pour les cibles « littérales » et les cibles liées au dernier mot d'une expression, ce qui confirme la nécessité de ne pas se fier entièrement à ce dernier type de cibles pour évaluer l'activation d'une interprétation littérale globale.

Un amorçage idiomatique significatif n'est observé que lors de la deuxième expérience. La relative faiblesse de l'effet (19 ms), qui n'est significatif que dans l'analyse par sujets, tient sans doute pour une part à ce que, les contextes de présentation étant neutres, le sens idio-

matique des expressions ne s'est trouvé que peu activé.

Cependant, on remarque également que le seul délai auquel l'amorçage idiomatique est significatif est aussi le seul pour lequel l'amorçage littéral (bien que descriptivement identique) n'est plus significatif. Cette inversion des patterns d'amorçage littéral et idiomatique semble indiquer l'existence d'un phénomène d'interférence (Titone & Connine, 1994, parleraient « d'inhibition »). Lorsque les deux interprétations sont activées, cette interférence provoquerait une diminution de leurs niveaux d'activation respectifs, sans pour autant les faire disparaître totalement (le fait que, dans les résultats présents, l'interprétation idiomatique soit momentanément « gagnante » peut s'expliquer par le fait que 13 des 15 expressions utilisées possédaient, selon les estimations recueillies auprès d'un ensemble de 20 étudiants, une interprétation figurée largement plus fréquente que leur interprétation littérale).

L'ensemble des résultats décrits plus haut est largement compatible avec un modèle de traitement des expressions idiomatiques proposé par Stock, Slack & Ortony (1993), dans lequel la représentation de chaque expression dans le lexique mental est associée à celle de chacun des mots qui la composent, de façon plus ou moins forte selon l'importance du mot dans l'expression. La présence d'un ou de plusieurs mots associés à une expression idiomatique déclenche la mise en oeuvre d'un processus destiné à évaluer la probabilité que cette expression se trouve réellement dans la phrase traitée. Si cette probabilité est suffisamment élevée, un nouveau processus débute, qui donne lieu à la construction du sens idiomatique. Dans le cas contraire, l'hypothèse de la présence d'une expression idiomatique est abandonnée.

Jusqu'ici assez semblable au modèle « de la configuration » de Cacciari & Tabossi (1988), dont il s'inspire d'ailleurs, ce modèle s'en distingue cependant en postulant que, pendant tout le temps où une éventuelle interprétation figurée est recherchée, l'analyse littérale n'est pas suspendue: elle se poursuit en parallèle, et, au cas où l'analyse idiomatique conduirait à un échec, peut aboutir sans qu'aucun retour en arrière ne soit nécessaire.

Dans ce cadre théorique, il est ainsi possible d'expliquer le fait que l'interprétation littérale globale des expressions idiomatiques soit dérivée (puisqu'une analyse syntaxico-sémantique est réalisée sur tous les énoncés traités), et que, bien que son degré d'activation soit temporairement réduit (Expérience 2), elle puisse être de nouveau disponible lorsque l'interprétation idiomatique n'est plus retenue comme valide (Expérience 3). De plus, ce modèle rend compte du fait que l'activation du sens idiomatique est, du moins lorsque les expressions sont présentées dans des contextes neutres, relativement plus lente que celle de l'interprétation littérale, puisqu'elle impliquerait un certain nombre d'opérations supplémentaires, comme par exemple, l'évaluation de différentes hypothèses.

3. BIBLIOGRAPHIE

Cacciari, C. & Tabossi, P. (1988) The comprehension of idioms. *Journal of Memory and Language*, 27, 668-683.

Colombo, L. (1993). The comprehension of ambiguous idioms in context. In C. Cacciari & P. Tabossi (Eds.), *Idioms: Processing, structure, and interpretation* (pp. 163-200). Hillsdale, NJ: Erlbaum.

Gibbs, R. W. (1980). Spilling the beans on understanding and memory for idioms in conversation. *Memory and Cognition*, 8, 149-156.

Gibbs, R. W., & Nayak, N. P. (1989). Psycholinguistic studies on the syntactic behavior of idioms. *Cognitive Psychology*, 21, 100-138.

Peterson, R. R., & Burgess, C. (1993). Syntactic and semantic processing during idiom comprehension: Neurolinguistic and psycholinguistic dissociations. In C. Cacciari & P. Tabossi (Eds.), *Idioms: Processing, structure, and interpretation* (pp. 229-247). Hillsdale, NJ: Erlbaum.

Rey, A., & Chantreau, S. (1991). *Dictionnaire des expressions et locutions*. Paris: Dictionnaires Le Robert.

Stock, O., Slack, J., & Ortony, A. (1993). Building castles in the air. Some computational and theoretical issues in idiom comprehension. In C. Cacciari & P. Tabossi (Eds.), *Idioms: Processing, structure, and interpretation* (pp. 229-247). Hillsdale, NJ: Erlbaum.

Swinney, D. A. (1981). Lexical processing during sentence comprehension: Effects of higher order constraints and implications for representation. In T. Myers, J. Laver, & J. Anderson (Eds.), *The Cognitive Representation of Speech* (pp. 201-209).Amsterdam: North-Holland.

Swinney, D.A., & Cutler, A. (1979). The access and processing of idiomatic expressions. *Journal of Verbal Learning and Verbal Behavior*, 18, 523-534.

Titone & Connine (1994). Comprehension of idiomatic expressions: Effects of predictability and literality. Journal of Experimental Psychology: Learning, Memory, and Cognition, 20, 1126-1138.

PLACE DU PHONOLOGIQUE ET DU VISUEL DANS LES ERREURS EN LECTURE ET EN ECRITURE

ANALYSE À PARTIR D'UN CORPUS PORTANT SUR DES NORMOLECTEURS DE CP

Liliane SPRENGER-CHAROLLES, URA 103, Université Paris V - René Descartes
UFR Linguistique Générale et Appliquée; 1 rue Cujas, 75005 Paris
Tél: (1) 40 46 29 75
&
Georges BOULAKIA, EA 333, Université Paris 7 - Denis Diderot
Laboratoire de phonétique de l'UFRL; 10 rue Charles V, 75004 Paris
Tél: (1) 44 78 34 59 - Fax: (1) 44 27 79 19 - georges.boulakia@linguist.jussieu.fr

ABSTRACT

The central hypothesis of this study was that phonological mediation plays a critical role in the early development of reading and spelling in French. Therefore, the phonological structure of items, as opposed to their visual characteristics, was expected to be a significant determinant of performance. This hypothesis was tested in a longitudinal study with a group of first graders who were administered a reading and a spelling task involving pseudowords of different syllabic structures. The first prediction was that there would be better performance on pseudowords with a simple structure (CVCVCV) as compared to pseudowords with a complex structure (CCVCVC or CVCCVC) and that errors on syllables with a complex structure would involve the deletion of codas or the simplification of complex onsets. Secondly, we predicted that errors would be consistent with a sonority hierarchy, for example, we expected more deletions of liquids than obstruents in clusters. Thirdly, we predicted that the principal phonological categories would be preserved in substitutions (for example, stops would be substituted by stops). The findings corroborated the predictions and confirmed the importance of phonological mediation in the initial stage of the acquisition of reading and spelling.

1. INTRODUCTION

Selon Orton (1937) et Boder (1973), les difficultés de lecture relèvent pour une large part de problèmes visuels qui se manifestent par des confusions entre lettres (de type p/b, b/d) et par des inversions de séquences (de type ble/bel). Plus généralement, dans pratiquement toutes les classifications des erreurs de lecture ou d'écriture, on répertorie ce type d'erreurs dans les erreurs visuelles (Aaron, 1994). Vellutino a fortement critiqué cette position dans son ouvrage de 1979 (voir également Fisher, Liberman & Shankweiler, 1977; Liberman, Shankweiler, Orlando, Harris & Bell-Berti, 1971) dans lequel il montrait que l'aspect linguistique du matériel n'était pas pris en compte dans ces classifications des erreurs.

L'objectif de notre étude est de réexaminer cette question. Pour ce faire, nous avons présenté à des enfants de CP une épreuve de lecture et une épreuve d'écriture de pseudomots dans laquelle nous avons contrôlé linguistiquement le matériel.

2. PRESENTATION DE L'ETUDE

Si les enfants sont sensibles aux caractéristiques phonologiques des items, leur structure syllabique devrait avoir une forte incidence sur les performances à l'écrit. Ainsi la structure syllabique CV est très prégnante puisqu'elle se retrouve dans toutes les langues et qu'il existe des langues qui n'ont que des syllabes de type CV. En conséquence, on peut prédire des performances supérieures pour les items qui présentent ce type de structure syllabique (CV) comparativement à des items contenant un groupe consonantique en attaque (CCV) ou une rime avec coda (CVC). On peut également prédire que les erreurs portant sur les items à structure syllabique complexe devraient être surtout des erreurs de suppression des codas (CVC→CV) ou de simplification des attaques complexes (CCV→CV). En outre, ces erreurs devraient respecter la hiérarchie de sonorité: les éléments les plus sonores devraient plus facilement s'assimiler aux voyelles quelle que soit leur place, après une consonne moins sonore dans une attaque complexe ou en coda intersyllabique avant une consonne moins sonore. Enfin, on peut supposer que les catégories phonologiques principales des items seront préservées. Ainsi, dans le cas des substitutions, les voyelles devraient être remplacées par des voyelles et les consonnes par des consonnes de la même catégorie phonologique: occlusive pour occlusive, fricative pour fricative et liquide pour liquide.

Ces principes généraux devraient s'appliquer dans toutes les langues avec des variations liées aux spécificités de chaque langue. Par exemple, en raison de la stabilité des voyelles en français qui gardent en général un timbre précis quelle que soit

leur position, on ne devrait observer que peu d'erreurs sur les voyelles.

Ces questions ont été examinées à partir d'épreuves de lecture à haute voix et d'écriture sous dictée de pseudomots de même longueur mais de différentes structures syllabiques: des CV/CV/CV, des CVC/CVC et de CCV/CVC (par exemple, *tibulo*, *tirbul*, *tribul*). Les résultats proviennent d'une étude longitudinale qui a porté sur 57 enfants francophones scolarisés en CP dans 20 classes différentes. Ces sujets, qui ont été vus en milieu et en fin d'année scolaire, ont un niveau cognitif moyen à supérieur; ils ne présentent aucun handicap majeur (langagier, sensori-moteur, sociologique) et n'ont pas de troubles psychologiques.

3. RESULTATS ET DISCUSSION

Dans cette étude, quatre hypothèses ont été évaluées. La première hypothèse est que, en raison de la prégnance de la division des syllabes en attaque/rime avec comme constituant de base une attaque simple et une rime sans coda, on devrait observer, d'une part, des performances supérieures pour les items de structure syllabique simple (CV) comparativement aux complexes (CCV OU CVC) et, d'autre part, des erreurs essentiellement sur les items à structure syllabique complexe de type suppression des codas ou simplification des attaques complexes. Les résultats corroborent cette hypothèse. En effet, les enfants lisent et écrivent mieux les items à structure syllabique simple (CVCVCV: par exemple, *tivulo*) que les complexes, qu'il s'agisse des CCV/CVC (par exemple, *trivul*) ou des CVC/CVC (par exemple, *tirvul*). L'analyse des erreurs a également permis de voir que pratiquement toutes les suppressions, en lecture comme en écriture, concernent la simplification d'une attaque complexe (*tivul* pour *trivul*) ou la disparition d'une coda (*tirvu* ou *tivul* pour *tirvul*); ces erreurs ont donc tendance à réduire la syllabe à ses éléments principaux (tableau 1).

La seconde hypothèse est que les erreurs devraient respecter les principes de sonorité. De fait, les résultats montrent que les éléments les plus sonores sont ceux qui sont les plus sujets à suppression (tableau 1) aussi bien quand ils sont situés après une consonne moins sonnante dans une attaque complexe (t[r]ibul) que quand ils sont dans une coda intersyllabique avant une consonne moins sonnante (ti[r]bul). Ces résultats indiquent que les suppressions sont induites par les propriétés phonologiques des consonnes et non par leur place.

La troisième hypothèse est que les catégories phonologiques principales des items devraient être préservées. Dans ce cas également les résultats observés sont conformes à ceux qui étaient attendus. Ainsi, l'analyse des substitutions (tableau 2) montre que les voyelles sont remplacées par les voyelles et les consonnes par des consonnes qui, en outre, appartiennent à la même catégorie phonologique: occlusive pour occlusive, fricative pour fricative, liquide pour liquide. Les résultats ont également permis de voir que, pour les occlusives, les substitutions dites visuelles entre p/b ou b/d ne sont pas plus nombreuses que celles portant sur leurs équivalents phonologiques t/d et p/t sauf pour la première session en lecture. En outre, toujours à l'exception de la première session en lecture, les substitutions concernant le voisement (p/b et t/d) sont plus nombreuses que celles portant sur le lieu d'articulation (b/d et p/t). Ce dernier résultat est conforme à ce qui est généralement relevé pour les erreurs en production orale (cf. par exemple, MacKay, 1992). De plus, les inversions de lettres (*tribul* vs *tirbul*), qui sont souvent répertoriées dans les erreurs visuelles, ont été essentiellement observées à l'intérieur de la structure syllabique (tableau 3). Ces erreurs peuvent donc également être phonologiques.

Le seul résultat contradictoire par rapport à l'hypothèse d'une nature principalement phonologique des erreurs est celui observé pour la première session en lecture. En effet, dans cette session pratiquement la moitié des substitutions portent sur b/d. Ce résultat indique que pour la lecture, à cette étape du développement, les enfants ont des difficultés de type visuel avec les lettres qui ne se différencient que sur l'axe gauche - droite, celui de l'orientation de l'écriture.

Enfin, la dernière hypothèse est que les principes phonologiques généraux doivent s'appliquer avec cependant des variations liées aux spécificités des langues. Ainsi, dans la mesure où les voyelles en français gardent généralement un timbre précis qu'elles soient ou non accentuées, on attendait peu d'erreurs à ce niveau. Les résultats corroborent cette prédiction (tableau 4). En effet, on observe plus de réponses correctes sur les voyelles que sur les consonnes sauf dans un cas: en lecture pour les CV/CV/CV. Cependant, même dans ce cas, les réponses correctes ne sont pas plus nombreuses sur les consonnes. De même, l'analyse des erreurs montre que les substitutions et les suppressions portent rarement sur les voyelles (tableaux 1 et 2).

Cet ensemble de résultats est intéressant dans la mesure où, en anglais, on observe systématiquement un plus grand nombre d'erreurs sur les voyelles que sur les consonnes en lecture comme en écriture (pour la lecture: Bryson & Werker, 1989; Werker, Bryson & Wassenberg, 1989; Fisher, Liberman & Shankweiler, 1977, Siegel & Faux, 1989; pour l'écriture: Stage & Wagner, 1992; pour lecture et écriture: Roeltgen, 1992; Treiman, 1993). Cette différence entre anglais et français permet de noter l'incidence de la langue cible sur les performances.

Dans l'ensemble, nos résultats montrent que ce sont les caractéristiques phonologiques des items - et non leurs caractéristiques visuelles - qui déterminent les performances des enfants au début de l'acquisition de la lecture/écriture. Ils répliquent ceux qui ont été obtenus en lecture/écriture de mots dans une autre étude longitudinale concernant également des enfants de CP (Sprenger-Charolles, 1993). En effet, dans cette étude, on a pu constater que les voyelles sont remplacées par des voyelles et les consonnes par des consonnes appartenant à une même catégorie phonologique (par exemple, *taple* pour *table*, *matame* pour *madame*). De plus, les éléments supprimés sont généralement situés dans des groupes consonantiques (par exemple, *butal* à la place de *brutal*) ou en fin de mot dans une coda simple (par exemple, *tiroi* pour *tiroir* ou *bou* pour *boule*). Ces erreurs préservent donc la structure syllabique des items, c'est-à-dire l'alternance consonne/voyelle, tout en la réduisant à une structure de type CV. Enfin, comme l'indiquent les exemples cités, les éléments supprimés sont surtout les consonnes les plus vocaliques ("r" ou "l"). Une autre tendance dégagée dans cette étude sur la lecture/écriture de mots, qui reproduit également celle observée pour les pseudomots, concerne le fait que les erreurs sont plus nombreuses sur les voyelles que sur les consonnes.

4. CONCLUSION

Nos résultats montrent les limites des catégorisations qui opposent erreurs phonologiques et visuelles et conduisent donc à s'interroger sur la pertinence des classifications des erreurs à partir desquelles ont été élaborées des typologies de dyslexiques (cf. Boder, 1973, Orton, 1937).

De plus, les résultats concernant les voyelles et les consonnes, qui indiquent qu'en français les premières sont plus stables que les secondes en lecture et en écriture - alors que l'inverse s'observe en anglais - montrent les limites des modèles élaborés et évalués principalement sur une seule langue.

5. BIBLIOGRAPHIE

Aaron, P.G. (1994). *Dyslexia and hyperlexia: Diagnosis and management of developmental reading disabilities.* Dordrecht/Boston/London: Kluwer Academic Press.

Boder, E. (1973). Developmental dyslexia: A diagnostic approach based on three atypical reading-spelling patterns. *Developmental Medicine and Child Neurology, 15*, 663-687.

Bryson, S.E. & Werker, J.F. (1989). Toward understanding the problem in severely disabled readers. Part I: Vowel errors. *Applied Psycholinguistics, 10*, 1-12.

Fisher, F.W., Liberman, I.Y. & Shankweiler, D. (1977). Reading reversals and developmental dyslexia: A further study. *Cortex, 14*, 496-510.

Liberman, I.Y., Shankweiler, D., Orlando, C., Harris, K.S. & Bell-Berti, L.B. (1971). Letter confusion and reversals of sequence in beginning readers: Implications for Orton's theory of developmental dyslexia. *Cortex, 7*, 127-142.

MacKay, D.G. (1992). Awareness and error detection: New theories and research paradigms. *Consciousness and Cognition, 1*, 199-225.

Orton, S.T. (1937). *Reading, writing and speech problems in children.* New York: Norton.

Roeltgen, D.P. (1992). Phonological error analysis, development and empirical evaluation. *Brain and Language, 43*, 190-229.

Siegel, L.S. & Faux, D. (1989). Acquisition of certain grapheme-phoneme correspondences in normally achieving disabled readers. *Reading and Writing: An Interdisciplinary Journal, 1*, 37-52.

Sprenger-Charolles, L. (1993). Procédures de traitement de l'information écrite utilisées par des lecteurs/scripteurs francophones en début d'apprentissage: examen à partir de l'analyse d'un corpus d'erreurs. *Etudes de Linguistique Appliquée, 91*, 70-83.

Stage, S.A. & Wagner, R.K. (1992). Development of young children's phonological and orthographic knowledge as revealed by their spellings. *Developmental Psychology, 28*, 287-296.

Treiman, R. (1993). *Beginning to spell: A study of first grade children.* New-York: Oxford University Press.

Vellutino, F.R. (1979). *Dyslexia: Theory and research.* Cambridge (Mass): M.I.T. Press.

Werker, J.F., Bryson, S.E. & Wassenberg, K. (1989). Toward understanding the problem in severely disabled readers. Part II: Consonant errors. *Applied Psycholinguistics, 10*, 13-30.

Tableau 1: Erreurs de suppression en lecture/écriture: pourcentage (effectif entre parenthèse)

SUPPRESSIONS Effectif:	Lecture Janvier (124)	Lecture Juin (031)	Ecriture Janvier (227)	Ecriture Juin (122)
Consonnes	**67%**	**81%**	**85%**	**77%**
Voyelles	**33%**	**19%**	**15%**	**23%**
Consonnes en attaque simple	**00%**	**00%**	**12%**	**04%**
Consonnes en attaque complexe	**63%**	**16%**	**38%**	**45%**
1ère syllabe:	10%	0%	10%	10%
1ère lettre: occlusive/fricative	90%	100%	90%	90%
2ème lettre: liquide	**37%**	**84%**	**50%**	**51%**
Consonnes en coda (liquide)	13%	38%	83%	73%
1ère syllabe: pré-cs (rt)	87%	62%	17%	27%
2ème syllabe: à la fin de l'item				

Tableau 2: Erreurs de substitution en lecture/écriture: pourcentage (effectif entre parenthèse)

SUBSTITUTIONS Effectif:	Lecture Janvier (092)	Lecture Juin (115)	Ecriture Janvier (105)	Ecriture Juin (097)
Voyelle/voyelle	**08%**	**09%**	**20%**	**14,5%**
Consonne/consonne	**92%**	**91%**	**77%**	**81,5%**
Voyelle/consonne	**00%**	**00%**	**03%**	**04,0%**
Consonnes de même catégorie phonétique:	**93%**	**97%**	**88%**	**86%**
Pour les occlusives:				
p/b	14%	23%	17%	20%
b/d	49%	24%	26%	23%
Total p/b, b/d:			**43%**	
t/d	**63%**	**47%**	47%	**43%**
p/t	37%	50%	06%	53%
Total t/d, p/t:	00%	00%	**53%**	04%
Divers (p/d, b/t)			**04%**	
	37%	**50%**		**57%**
	00%	**02%**		**00%**

Tableau 3: Erreurs d'inversion de séquence en lecture/écriture: pourcentage (effectif entre parenthèse)

INVERSIONS Effectif:	Lecture Janvier (026)	Lecture Juin (33)	Ecriture Janvier (13)	Ecriture Juin (14)
Dans la syllabe	92%	89%	100%	93%
Hors de la syllabe	08%	11%	0%	07%

Tableau 4: Réponses correctes pour les consonnes et voyelles: moyenne (écart-type entre parenthèse)

	Ensemble des items		CVC/CCV		CCV/CVC		CV/CV/CV	
	consonnes /88	voyelless /56	consonnes /32	voyelles /16	consonnes /32	voyelles /16	consonnes /24	voyelles /24
LECTURE								
Janvier	60.0 (27.9)	69.4 (28.6)	55.8 (29.2)	68.9 (29.3)	57.7 (30.9)	69.2 (33.5)	68.7 29.3)	69.8 (29.8)
Juin	84.9 (19.8)	88.6 (19.0)	82.9 (23.0)	87.0 (20.1)	85.1 (20.4)	89.9 (20.7)	87.2 (19.6)	88.7 (20.7)
ECRITURE								
Janvier	72.8 (25.7)	86.1 (16.4)	70.3 (26.6)	87.4 (16.4)	69.3 (26.9)	83.4 (20.3)	81.0 (25.8)	87.1 (16.6)
Juin	89.8 (15.7)	94.7 (8.2)	89.3 (16.8)	93.5 (9.52)	88.9 (17.2)	94.2 (10.6)	91.7 (16.7)	95.8 (7.91)

PREDICTION DES SYSTEMES VOCALIQUE PAR APPROCHE DEDUCTIVE

René CARRE

Département Signal, Unité Associée au CNRS, ENST - 46, rue Barrault - 75634 Paris cedex 13

Tél.: 45817190 - Fax: 45887935 - e-mail: carre@sig.enst.fr

ABSTRACT

A deductive approach is developed to predict vocalic systems. First, vowels are proposed from an efficient and simple use of acoustic tube. Then, the maximum acoustic dispersion criterion is applied to classify the vowels.

1. INTRODUCTION

Les systèmes vocaliques font l'objet de nombreuses recherches. Il s'agit d'abord d'accumuler le maximum de données sur les langues du monde (Maddieson, 1984), (Crothers, 1978), d'étudier le statut phonologique de tel ou tel son d'une langue, d'étiqueter et de classer ces voyelles (Vallée, 1994). Les systèmes comportent statistiquement des régularités que l'on essaie d'expliquer. Que nous disent aujourd'hui les données?

- que les voyelles des systèmes à 3, 4, 5, ... voyelles sont statistiquement /a, i, u/, /a, i, u, ɛ/, /a, i, u, ɛ, ɔ/,... ou bien /a, i, u/, /a, i, u, ɨ/, /a, i, u, ɨ, ɛ/,... Les systèmes se divisent donc en deux grandes classes selon qu'ils intègrent ou non la voyelle centrale /ɨ/ et les systèmes sont obtenus par ajout successif d'une nouvelle voyelle (Crothers, 1978).
- que les voyelles peuvent être classées en voyelles: orales labialisées ou non, nasales, ATR, rétroflexes, diphtongues... et en fonction de leur complexité, (Carré, 1994), (Carré, et al., 1995). Les premières voyelles des systèmes ne font appel qu'à des distinctions de place puis de degré de constriction de la langue, puis au geste distinctif labial, nasal,...
- que les systèmes peuvent être équilibrés ou non (Maddieson, 1984), avec une tendance à exploiter de manière maximale les traits les décrivant (Ohala, 1980).

Toutes ces constatations ne sont pas explicatives. Aussi, pour éviter des explications entachées de circularité, des approches à partir de critères extérieurs aux observations ont été développées. Il s'agit d'approches dites 'substance-based' (Liljencrants & Lindblom, 1972) ou déductives. Pour expliquer les systèmes vocaliques, on a exploité, coté perception, un critère de dispersion acoustique (Liljencrants & Lindblom, 1972), (Lindblom, 1986) appliqué dans l'espace vocalique. Coté production, Stevens (Stevens, 1972), utilisant un critère de stabilité de la relation articulatoire-acoustique (théorie quantique), propose que les voyelles correspondent aux zones de stabilité. Quant à ten Bosch (ten Bosch, 1991), il introduit un critère de minimum d'effort. Plus récemment, on a proposé d'exploiter et le critère dispersion avec un critère de focalisation c'est-à-dire de proximité de deux formants consécutifs (Vallée, 1994). Ces approches intéressantes ne permettent pas de donner des explications complètement satisfaisantes. Le critère de dispersion, d'une part, conduit à déplacer les positions des voyelles d'un système lors de l'ajout d'une nouvelle voyelle: on a donc toujours des ensembles 'équilibrés'; d'autre part, il fait apparaître plus de voyelles sur l'axe /iu/ que ce qui est observé dans les données au détriment des voyelles centrales. Par ailleurs, la prise en compte de la focalisation paraît être une explication ad'hoc qui convient bien pour la production de certains /y/ mais pas pour d'autres voyelles labialisées comme /ø, œ/. On peut d'ailleurs dire que la stabilité quantique n'est pas une caractéristique intrinsèque des voyelles. Si les voyelles /a, i, u/ sont stables, c'est parce qu'elles se situent aux extrémités des possibilités acoustiques du conduit vocal. Rappelons que la 4ème voyelle la plus courante, /ɛ/, est particulièrement instable.

En conclusion, nous pensons que le rôle de l'appareil vocal dans sa capacité à produire simplement et efficacement des successions de

sons n'est pas suffisamment étudié et exploité pour la prédiction des systèmes vocaliques. Or, comme nous l'avons dit, les passages de voyelle à voyelle, dans le cas des systèmes comprenant jusqu'à 7, 8 voyelles, ne font appel qu'à un seul geste distinctif. On peut donc les classer selon une caractéristique liée à la production. Si le cerveau a une grande capacité à traiter le signal et à commander l'appareil vocal, il doit savoir exploiter au mieux, avec le minimum d'effort et de complexité, les caractéristiques de cet appareil vocal. Par ailleurs, si l'on suppose que l'appareil de production est « optimal » au sens de la théorie de la communication, les caractéristiques de cet appareil doivent répondre aux critères suivants:

- critère de maximisation du contraste acoustique des voyelles représentées par leurs fréquences de formants (pour permettre un rapport signal-sur-bruit maximum),
- critère de minimum d'effort: efficacité de la relation articulatoire-acoustique incluant la monotonie de la relation et son orthogonalité,
- critère de simplicité: commandes simples (rectilignes?) et en nombre réduit.

Le critère de stabilité de la relation (aspects quantiques) est ici rejeté car contraire au critère de minimum d'effort et donc d'efficacité. On approche seulement les zones extrêmes de stabilité qui se trouvent sur les bords du triangle vocalique (et, en particulier, aux angles).

On va maintenant développer une approche déductive à partir du critère de contraste acoustique maximum.

2. CONTRASTE ACOUSTIQUE MAXIMUM: VERS LE MODELE DRM

On a montré par ailleurs (Carré, et al., 1995) que l'application d'un critère de contraste acoustique maximal sur un tube acoustique de forme quelconque permet de retrouver le triangle des voyelles ainsi que des formes du tube acoustique correspondant aux extrémités de ce triangle. Ces formes sont celles, simplifiées, du conduit vocal lors de la production des voyelles /a, i, u/. La figure 1 montre l'évolution automatique de la forme du tube acoustique à partir de la voyelle /i/ suivant un critère d'accroissement de F_1. On a imposé une cavité du larynx.

Les conclusions de cette étude sont les suivantes:

- obtention de configurations à contraste acoustique maximal (selon le critère de départ) correspondant aux 3 voyelles /a, i, u/ du triangle vocalique,

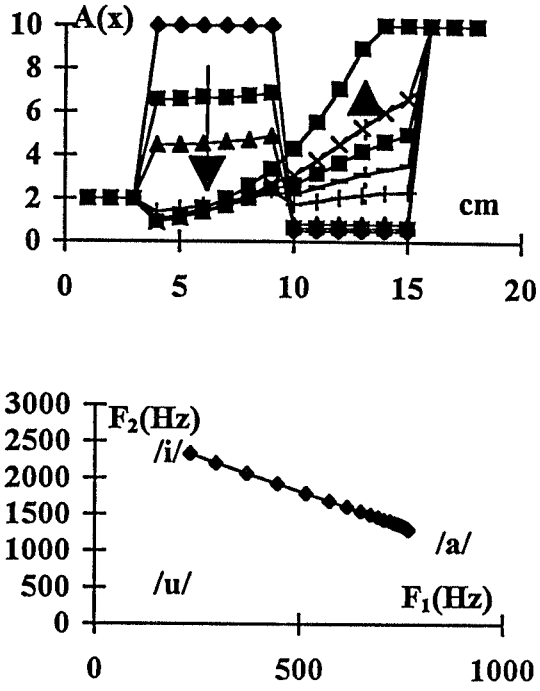


Figure 1: Evolution de la forme du tube acoustique à partir d'une configuration correspondant à la voyelle /i/ ainsi que l'évolution de la trajectoire formantique correspondante dans le plan F_1-F_2.

- commande avec minimum d'effort (conséquence de l'utilisation de la fonction de sensibilité), donc efficace, avec monotonie (liée à l'augmentation du contraste acoustique), et pseudo-orthogonalité,

- commandes (ou gestes) de déformation simples, en nombre limité (1 seule, rectiligne, dans le cas de la figure 1 réalisant la constriction arrière avec mise en place automatique, comme chez l'homme, d'une cavité avant).

On retrouve donc dans un même dispositif (le tube acoustique), toutes les caractéristiques qui nous paraissaient utiles au bon fonctionnement d'un émetteur de signaux de communication. On montre aussi que l'on retrouve les deux structures (fermé-fermé et fermé-ouvert) du modèle DRM (Carré & Mrayati, 1992).

3. PRODUCTION ET REPRESENTATION DES VOYELLES

Avec le dispositif précédent, structuré selon le modèle DRM, nous allons développer une approche systémique topologique permettant de produire simplement des sons (formes simples dont le passage de l'une à l'autre est simple à effectuer et donc à représenter). Le modèle DRM, exploitant les régions acoustiquement sensibles (entre des stabilités quantiques!), permet une quantification du tube en régions et donc une classification simple des sons produits selon ces régions (constriction avant, arrière et centrale par exemple). Toujours par soucis de simplicité, le nombre de degrés de constriction va être réduit. Par ailleurs, un nouveau degré de constriction devra être totalement exploité pour former un ensemble topologiquement 'équilibré' (exploitation maximale d'une nouvelle possibilité de production). On a, par ailleurs (Carré & Mrayati, 1995), étudié le passage entre les deux configurations types du DRM (le modèle fermé-ouvert et le modèle fermé-fermé).

Finalement, les voyelles produites avec un maximum de deux commandes: geste de déplacement de la constriction (deux degrés de constriction de 0.5 et 1.3cm^2 plus le neutre de 4cm^2) et geste de labialité (un degré de constriction), sont représentées figure 2. On note que la position de la constriction est topologiquement simplement positionnée le long du conduit vocal : aux extrémités, au centre, etc... La division de la partie arrière en tiers vient d'un examen des caractéristiques acoustiques obtenues, la voyelle /ɔ/ étant plus ou moins acoustiquement équidistante de /ɑ/ et de /u/. La voyelle /ɛ/ est classée ici soit comme constriction arrière, soit comme constriction avant. Les passages d'une voyelle à une autre sont phonologiquement représentés par des 'gestes intentionnels' simples (Carré, et al., 1995).

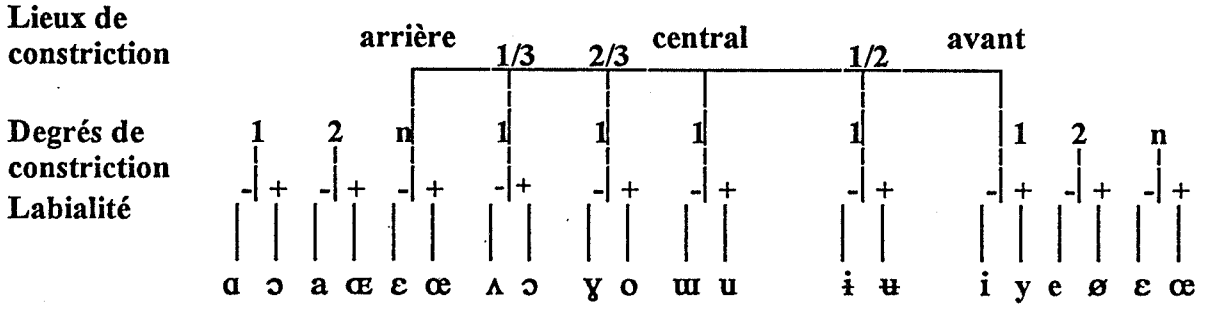


Figure 2: Voyelles obtenues à partir du modèle DRM. Constrictions 1, 2, n, respectivement: 0.5, 1.3, 4 cm^2.

4. PREDICTION DES SYSTEMES VOCALIQUES

Pour construire des systèmes vocaliques à 3, 4, 5,.. voyelles, nous allons successivement choisir parmi les voyelles obtenues précédemment celles qui répondent au critère de 'dispersion' acoustique maximale. Les 3 premières voyelles sont /ɑ (ou /a/), i, u/, correspondant au contraste acoustique et topologique (constriction arrière, avant puis centrale) maximum. La production du /u/ ne fait pas ici appel à un geste labial distinctif.

Puis, deux choix sont possibles selon que l'on retient ou non (voir (Crothers, 1978)) la voyelle centrale /ɨ/ (milieu de la trajectoire formantique /iu/ et constriction au milieu de la distance avant-centrale). Cette voyelle est très instable car les gestes de constriction et de labialité ne produisent pas des trajectoires acoustiques orthogonales comme c'est le cas pour /ɛ/ (milieu de la trajectoire formantique /ɑi/ et tube neutre). Par cette procédure, on obtient la progression (1, 2, 3,...) présentée figure 3 pour des voyelles qui ne font pas appel au geste distinctif de labialité. Dans la classe des systèmes sans /ɨ/, le système à 5 voyelles est équilibré; c'est aussi le plus courant. Avec la voyelle centrale /ɨ/, le système est équilibré avec 6 voyelles et c'est aussi le plus courant. L'introduction du geste distinctif de labialité introduira automatiquement les

voyelles labialisées /œ/, puis /y/, puis /ø/, en suivant le critère de dispersion acoustique. Dans la classe des systèmes avec /ɨ/, il faudrait examiner plus en détail le système avec 7 voyelles (apparition de /e/, /o/, /ə/).

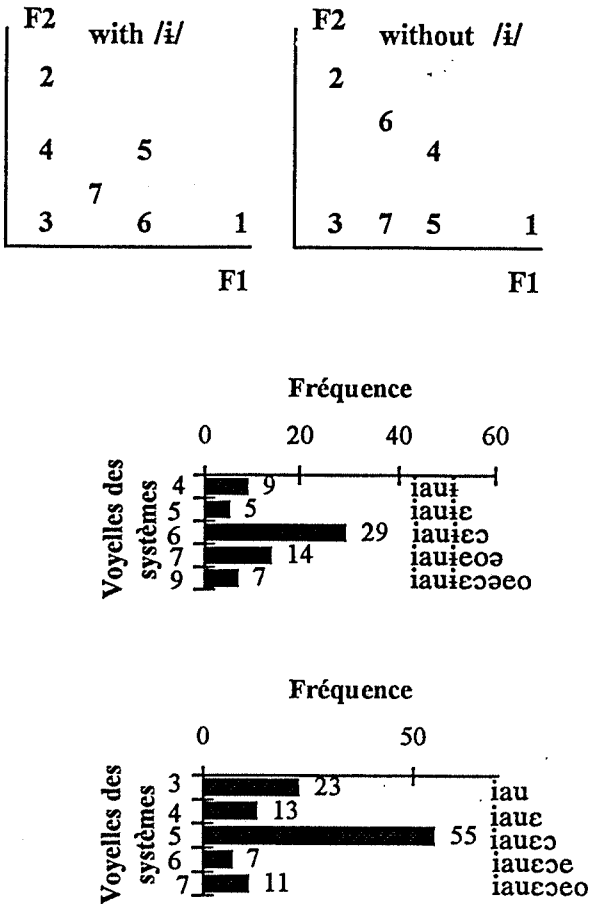


Figure 3: Choix des voyelles successives avec /ɨ/ ou non (en haut). Statistiques sur les systèmes vocaliques avec /ɨ/ ou non (en bas) (d'après (Crothers, 1978)).

L'approche s'appuyant sur les capacités d'un dispositif acoustique à produire simplement et efficacement des sons, éléments de code, répond aux principales critiques formulées contre les approches évoquées précédemment. La pseudo-orthogonalité de la relation articulatoire-acoustique mise en évidence dans le cas d'un tube divisé en régions distinctives (DRM), explique la bonne correspondance et la complémentarité, simples à exploiter, des deux démarches systémiques: topologique d'une part et acoustique d'autre part.

Remerciements. Cette recherche a fait l'objet d'un soutien du Programme Science de l'Union Européenne.

5. BIBLIOGRAPHIE

Carré, R. (1994) "Speaker" and "Speech" Characteristics: a deductive Approach. *Phonetica, 51, 7-16.*

Carré, R., Bourdeau, M. & Tubach, J.P. (1995) Vowel-vowel production: the distinctive region model (DRM) and vocalic harmony. *Phonetica, 52, 205-214.*

Carré, R., Lindblom, B. & MacNeilage, P. (1995) Rôle de l'acoustique dans l'évolution du conduit vocal humain. *CR Acad. Sci. Paris, t. 30, série IIb, 471-476.*

Carré, R. & Mrayati, M. (1992) Distinctive regions in acoustic tubes. Speech production modeling. *J. d'Acoustique, 5, 141-159.*

Carré, R. & Mrayati, M. (1995) Vowel transitions, vowel systems, and the Distinctive Region Model, in *Levels in Speech Communication: Relations and Interactions* (C. Sorin, J. Mariani, H. Méloni and J. Schoetgen, Eds.), Amsterdam: Elsevier.

Crothers, J. (1978) Typology and universals of vowel systems, in *Universals of human language. Vol. 2: Phonology* (J. H. Greenberg, C. A. Ferguson and E. A. Moravcsik, Eds.), Stanford: Stanford University Press.

Liljencrants, J. & Lindblom, B. (1972) Numerical simulation of vowel quality systems: The role of perceptual contrast. *Language, 48, 839-862.*

Lindblom, B. (1986) Phonetic Universal in Vowel Systems, in *Experimental Phonology* (J. J. Ohala and J. J. Jaeger, Eds.), Orlando: Academic Press.

Maddieson, I. (1984) *Patterns of sounds.* Cambridge: Cambridge University Press.

Ohala, J. (1980) Moderator introduction to symposium on phonetic universals in phonological systems and their explanations. *Proc. of the 9th ICPhS,* Copenhagen.

Stevens, K.N. (1972) The quantal nature of speech: evidence from articulatory-acoustic data, in *Human Communication: a unified view* (E. E. David and P. B. Denes, Eds.), New York: Mac Graw-Hill.

ten Bosch, L. (1991) On the structure of vowel systems; aspects of an extended vowel model using effort and contrast.Dissertation Thesis. Amsterdam.

Vallée, N. (1994) Systèmes vocaliques : de la typologie aux prédictions.Thèse de Doctorat en Sciences du Langage. Grenoble.

QUELQUES ASPECTS ACOUSTIQUES DE LA PRODUCTION DES OCCLUSIVES DU CORÉEN ET DU FRANÇAIS: ANALYSE COMPARATIVE

(VOT, durée de la consonne, durée de la voyelle précédente)

Hyeon-Zoo KIM, André BOTHOREL

Dankook University of South Korea Tél.: + 343 44 5161, 343 68 2450 - Fax: + 343 44 5161

Institut de Phonétique de Strasbourg - USHS 22 rue Descartes, 67084 Strasbourg Cedex, France

Tél.: 88 41 73 69, 88 41 73 00 - Fax: 88 41 73 69

ABSTRACT

This is a comparative study of the production of French and Korean consonants plosives used in continuous speech, using experimental analysis.

The Korean consonantal system has three unvoiced plosives:

	bilabials	alveodentals	palatals	velars
lenis	p	t	c	k
aspirated	p^h	t^h	c^h	k^h
glottalized	p‘	t‘	c‘	k‘

It is admitted for French that there are two series of plosives:

unvoiced	p	t	k (fortis)
voiced	b	d	g (lenis)

They differ as to their voicing status (voiced/ unvoiced) and their degree of tension (lenis/fortis).

The main question addressed here is, how can one explain the difference in behaviour for apparently the same elements that are in a similar context? The answer to this question is discussed in relation to the notion of the phonetic context and its influences. An additional theoretical notion should also be taken into account; that of a system, a decisive factor that can not be dissociated from the notion of "phonological constraints". A series of experiments are presented, in which identical stimuli are used to examine similar cue-trading relations in the perception of the voicing contrast in word stops in Korean and French. Predicted cross-linguistic differences are found inthe basis category boundary and in the case of cue trading between VOT and aspiration.

1. INTRODUCTION

L'analyse des matériaux de la présente étude nous a permis, d'une part, d'apporter quelques compléments aux résultats d'autres études et, d'autre part, d'en d'obtenir de nouveaux particulièrement intéressants en ce qui concerne le VOT. Nous rappelons que le VOT est un retard de la vibration des cordes vocales entre l'explosion de la consonne et le début de la voyelle subséquente (Abramson et Lisker, 1972). Cette étude nous a fourni l'occasion de pouvoir discuter la plupart des théories émises jusqu'à cette date à propos du VOT coréen et du VOT français, s'agissant des consonnes que nous avons retenues, à savoir les occlusives qui, en français et en coréen, présentent des différences assez importantes. En effet, le coréen comporte trois séries d'occlusives sourdes [p t k c - p^h t^h k^h c^h - p' t' k' c'] et le français, deux séries d'occlusives qui s'opposent par la sonorité [p t k - b d g].

2. PRÉLIMINAIRES

2.1. Méthode

Pour bien mesurer la durée en millisecondes, nous avons utilisé l'ordinateur Macintosh. Les systèmes d'enregistrement décrits utilisent à bon escient les capacités du Macintosh en matière de gestion du son. A l'instar des produits décrits, le système "Audiomedia" conçu par Digidesign (aujourd'hui remplacé par Audiomedia II.) consiste en un ensemble matériel/logiciel permettant l'enregistrement et l'édition de son numérique sur ordinateur. "Audiomedia" enregistre directement sur le disque dur, à partir duquel se fait également la lecture. Les processeurs de la carte "Audiomedia" améliorent énormément les capacités audios de l'ordinateur, tout en lui permettant de gérer les tâches exigeantes en temps de calcul. A partir des données fournies par cette technique, nous avons retenu les paramètres suivants:

- VOT
- Durée totale de la consonne
- Durée de la voyelle précédente

2.1.1. Choix du corpus

Nous avons constitué un corpus, dont les items sont du type V1CV2 (voyelle-consonne-voyelle), où V1 représente la voyelle [a], tandis que V2 représente [a] [i] [u], les extrémités du triangle vocalique. C représente les occlusives.

Pour étudier les consonnes occlusives du français, nous nous appuyons sur un corpus

Tableau I. VOT, Durée totale de la consonne, Durée de la voyelle précédente [les trois groupes : coréen (C), français (F), français parlé par des coréens (CF)]

	VOT			Con.			Voy.		
	Bila.	Alvé.	Vél.	Bila.	Alvé.	Vél.	Bila.	Alvé.	Vél.
Fai. (C)	0	0	0	53	46	48	44	28	93
As. (C)	38	49	36	114	123	125	101	68	95
Glo. (C)	14	14	21	140	109	132	93	108	97
Acc. (F)	17	27	31	109	114	108	84	103	84
Inacc. (F)	13	19	21	99	101	106	86	74	75
Acc.(CF)	20	16	30	189	165	171	92	116	105
Inac.(CF)	19	14	19	129	137	103	89	101	87

constitué de 36 phrases pour les trois locuteurs français et les trois locuteurs coréens pour bien comparer les deux groupes de locuteurs.

Pour les occlusives coréennes, nous nous sommes appuyés sur un corpus constitué de 27 phrases pour les trois locuteurs coréens. Les phrases contiennent entre 4 et 7 syllabes en coréen, et il y a trois catégories : faibles, aspirées et glottalisées.

2.1.2. Choix des sujets

Les trois sujets coréens sont tous nés à Séoul. Ils ont tous accompli des études supérieures à Séoul. Leur âge se situe entre 30 et 34 ans. Tous ont le coréen pour langue maternelle. Les trois sujets français sont de langue maternelle française et sans accent régional, de sexe masculin, et dont l'âge varie de 21 ans à 27 ans.

2.2. RÉSULTATS

Dans cet article, pour comparer les occlusives sourdes [p, t, k] en français et les trois séries, faibles, aspirées, glottalisées coréennes, nous avons choisi les consonnes occlusives en position intervocalique [a]; cette position est, en effet, celle qui peut nous permettre une comparaison qui répond le mieux à la condition de toutes choses égales par ailleurs. La voyelle [a], de plus, marque le moins les consonnes qui l'entourent, à la différence de [i] qui a tendance à palataliser, surtout en français, la consonne qui précède.

Si l'on considère les valeurs du VOT du coréen parlant français, on constate une sorte de neutralisation qui fait que les valeurs relevées se situent dans une position intermédiaire entre celles qui caractérisent les aspirées du coréen et celles qui caractérisent les faibles (absence de VOT). Est-ce à dire que ces occlusives sont glottalisées dans la mesure où les valeurs sont très proches? Il ne nous semble pas puisqu'il y a une dispersion plus importante dans le résultat des occlusives dites par les Coréens parlant français par rapport aux glottalisées coréennes qui sont nettement moins dispersées.

D'après nos analyses dans ces différents aspects, nous essayerons ainsi de définir le rôle du VOICE ONSET TIME dans la distinction des trois séries d'occlusives coréennes (faibles : [p t k], aspirées : [p^h t^h k^h], glottalisées : [p' t' k']) et des deux positions d'occlusives sourdes françaises ([p t k] en position accentuée et inaccentuée) à partir de nos propres résultats expérimentaux.

Le VOT est, comme nous l'avons vu, l'élément déterminant dans la distinction entre les trois catégories de consonnes occlusives du coréen, puisque les durées tant absolues que relatives des occlusives aspirées se différencient de façon très significative de celles des non-aspirées.

En coréen, le VOT est donc, parmi les indices qui permettent de distinguer les trois catégories d'occlusives sourdes coréennes, celui qui est le plus important dans la distinction entre les occlusives aspirées et les occlusives glottalisées en position intervocalique. Le VOT du français est supérieur à celui des consonnes glottalisées du coréen. Chez les Coréens parlant français, la valeur du VOT n'est pas très élevée et elle correspond à celle des résultats obtenus pour les locuteurs français, comme elle correspond à la valeur obtenue pour les consonnes occlusives glottalisées du coréen, à peu de choses près.

La durée totale de la consonne du coréen permet de différencier les occlusives aspirées, les occlusives glottalisées et les occlusives faibles coréennes [p t k] en position intervocalique ; c'est toujours l'occlusive sonorisée qui présente la durée totale la plus restreinte.

Soulignons que la réalisation sur le plan de la durée totale des occlusives dites par les Coréens parlant français paraît presque toujours plus longue que celle des aspirées et des glottalisées coréennes et a fortiori plus longue que celle des occlusives du français.

Cette particularité ne serait-elle pas due aux faits que les locuteurs coréens parlant français auraient mis plus d'énergie et d'attention à la prononciation des phrases de la langue étrangère? Ce serait à considérer comme une sorte d'hyper-correction due aux circonstances particulières de l'enregistrement.

En ce qui concerne la durée de la voyelle précédente, nous devons reconnaître qu'elle

n'est pas un indice suffisant pour distinguer entre elles les occlusives du coréen, du moins dans le cas des aspirées et des glottalisées.

La durée de la voyelle précédente est nettement plus brève devant les occlusives faibles vélaires par rapport aux occlusives faibles bilabiales et alvéodentales.

Dans le cas des occlusives accentuées/inaccentuées du français, ainsi que dans le cas des Coréens parlant français, la durée de la voyelle précédente, étant pratiquement identique pour les exemples analysés, ne constitue pas un indice de différenciation.

Nous croyons qu'il est justifié de proposer la conclusion suivante : pour pouvoir aboutir à une description satisfaisante des trois séries d'occlusives coréennes [p t k], [p^h t^h k^h], [p' t' k'] et des deux séries d'occlusives françaises [p t k], [b d g] en position accentuée et inaccentuée, il faut prendre parti pour l'indépendance des traits de sonorité, de tension et d'aspiration, en accord avec les conclusions de Kim H.G. (1987). En conséquence, les occlusives sourdes coréennes ont des traits indépendants de force articulatoire et d'aspiration dont les indices sont différents en fonction des positions. Cependant, dans ce système qui ne connaît pas l'opposition de sonorité (sourde/sonore), nous proposons que la force articulatoire est l'un des traits distinctifs, au même titre que le trait d'aspiration.

3. DISCUSSION

Il nous semble utile de comparer nos résultats avec ceux de quelques travaux récents sur le VOT, en particulier les travaux de Flege.

Comme nous en avons déjà parlé dans une première partie, nous avons souvent observé que les occlusives faibles [p t k] en coréen peuvent phonétiquement être réalisées sonores en position intervocalique.

Quelques travaux sur la production de l'occlusive en position initiale chez les bilingues-anglais/espagnol (Williams, 1977) ont suggéré que les résultats de la production de l'occlusive pour les bilingues sont conformes :

- aux résultats pour les monolingues
- à une tendance pour les bilingues au prévoisement [b] en anglais.

Des postulats concernant uniquement les bilingues qui ont appris leur langue secondaire plus tard, peuvent établir la séparation "switch" des catégories phonétiques pour les deux langues (Flege, 1991). Dans cette étude, il apparaît qu'il y a une alternance (code switching) dans la production de toutes les catégories de locuteurs bilingues, bien que les sujets bilingues ne puissent pas exprimer complètement les valeurs monolingues dans leur langue secondaire. Le changement, dans la production du VOT de bilingues, est plus important que celui des résultats obtenus par Caramazza (French-English bilinguals) (1974). Il y a aussi une tendance très forte, comme dans la recherche de Williams (1977), pour le prévoisement /b/ de bilingues en anglais.

Caramazza & Yoni-Komshian (1974) ont conclu :

"... VOT is a sufficient phonological cue for the distinction of the homorganic stop consonant pairs in French (France). We have also proposed an explanation for the observed difference between French and Canadian French based on a linguistic change hypothesis." (p. 245).

Flege (1991), d'autre part, a supposé que la séparation complète de sons des inventaires phonétiques est tout au moins possible pour les personnes qui apprennent plus tôt une L2.

Ces diverses études ont montré que beaucoup d'adultes qui apprennent l'anglais en L2 produisent les [p t k] en anglais avec des valeurs du VOT significativement plus brèves que celles des Anglais monolingues mais, cependant, avec des valeurs du VOT plus longues que les sujets d'origine monolingue de L1 (Flege & Port, 1981; Port & Mitler, 1980; Flege, 1987).

Dans l'article de Flege (1987), il apparaît que les sujets les plus avancés dans leur maîtrise de l'anglais ont produit un [t] néerlandais avec des valeurs du VOT plus brèves que les sujets moins avancés; ce qui suggère qu'ils ont créé une nouvelle catégorie pour le [t] anglais.

Ce résultat concorde avec nos résultats. C'est-à-dire que, en coréen, le VOT est parmi les indices qui permettent de distinguer les trois séries d'occlusives sourdes coréennes, celui qui est le plus important dans la distinction entre les occlusives aspirées et les occlusives glottalisées en position intervocalique. Le VOT du français est supérieur à celui des consonnes glottalisées du coréen. Chez les Coréens parlant français (acc. : 22 ms; inacc. : 17 ms), la valeur du VOT n'est pas très élevée et elle correspond à celle des résultats obtenus pour les locuteurs français (acc. : 25 ms; inacc. : 18 ms), comme elle correspond sensiblement à la valeur obtenue pour les consonnes occlusives glottalisées du coréen (16 ms). Mais la valeur du VOT des occlusives glottalisées du coréen est un peu plus brève que celle du VOT du français parlé par les Coréens.

Flege a étudié le degré de dépendance du système phonétique de la langue maternelle et de la langue secondaire par l'examen détaillé de la production du /t/ en espagnol et en anglais par deux groupes différents : "early learners" et "late learners" d'espagnol.

Nos sujets coréens parlant français ont commencé à apprendre la langue française à l'âge de 20 ans (cela fait 10 ans), mais ne maîtrisent cette langue que depuis 4 ou 5 ans (depuis leur arrivée en France).

Comme nous l'avons vu précédemment, les sujets coréens ont différencié les 3 séries d'occlusives coréennes (faibles, aspirées et glottalisées). Les valeurs du VOT sont, en moyenne, nulles pour les faibles, de 40,9 ms pour les aspirées et de 16,3 ms pour les glottalisées. Chez les Coréens parlant français, la valeur du VOT de /p,t,k/ en français est, en moyenne, de 21,9 ms en position accentuée. Ce résultat est proche de celui des Français.

Flege (1987) indique que

"adults are capable of learning to produce new phones in an L2, and of modifying their previously established patterns of articulation when producing similar L2 phones. It appears that the mechanism of equivalence classification leads them to identify acoustically different phones in L1 and L2 as belonging to the same category. This may ultimately prevent them from producing similar but now new phones authentically," (p.62).

4. CONCLUSION

Nous pouvons donc conclure que les Coréens parlant français réalisent une autre forme d'occlusive. Il s'agit là d'une nouvelle catégorie d'occlusives sourdes.

Car les valeurs du VOT des occlusives sourdes [p t k] produites par les sujets coréens parlant français et maîtrisant la langue sont proches des valeurs du VOT du français. Mais il semblerait, d'après l'examen de la durée totale, que ni la consonne glottalisée du coréen, ni la consonne sourde du français ne soient utilisées par les Coréens parlant français et, au regard du VOT, qu'ils n'utilisent pas non plus la consonne aspirée, ce qui nous a confortés dans l'idée de l'emploi d'une catégorie nouvelle d'occlusive sourde.

Ceci est en accord avec les résultats obtenus par Flege dans ses études concernant l'apprentissage de l'anglais par des sujets néerlandais. En effet, il apparaît que les sujets ayant déjà une bonne maîtrise de l'anglais réalisent un [t] néerlandais avec des valeurs de VOT plus brèves que les sujets moins avancés. Cette observation a conduit Flege à supposer que les sujets créent une nouvelle catégorie pour le [t] anglais.

Dans un article de 1987, Flege indique que les adultes sont capables d'apprendre à réaliser des phones nouveaux dans une langue seconde, et à modifier leurs modèles articulatoires préalablement établis lorsqu'ils produisent des phones identiques de la langue seconde. Par conséquent, il apparaît que les sujets ont tendance à considérer sur le plan acoustique des phones différents dans la langue maternelle et dans la langue seconde comme appartenant à la même catégorie. Finalement, cela pourrait leur permettre de réaliser des phones non pas identiques mais véritablement nouveaux.

Cette constatation ouvre des perspectives intéressantes dans l'analyse des interférences entre deux langues telles qu'on les relève chez les bilingues ; la question qui reste posée est celle de savoir dans quelle mesure la didactique, en tant que science, peut prendre en compte cette notion de catégorie phonique nouvelle ?

5. BIBLIOGRAPHIE

Abramson A.S., Lisker L. (1971) Voice Timing in Korean Stops, *Haskins Laboratories Status Report on Speech Research, SR 27,* 179-184

Abramson A.S., Lisker L. (1972) Laryngeal behavior, the speech signal and phonological simplicity, *Actes du Xème Congr. Int. des Linguistes VI,* 123-129

Caramazza A., Yeni-Komshian G.H. (1974) Voice onset time in two French dialects, *Journal of Phonetics, 2,* 239-245

Flege J. (1987) The production of *"new"* and *"similar"* phones in a foreign language: evidence for the effect of equivalence classification, *Journal of Phonetics, 15,* 47-65

Flege J. (1991) Age of learning affects the authenticity of voice-onset time (V.O.T.) in stop consonants produced in a second language, *Journal Acoust. Soc. Am., 89 (1),* 395-411

Flege J., Port R. (1981) Cross-language phonetic interference: Arabic to English, *Language and Speech, 24,* 125-146

Flege J., Eefting W. (1988) Imitation of a VOT continuum by native speakers of English and Spanish: Evidence for phonetic category formation, *Journal Acoust. Soc. Am., 83,* 729-740

Kim H.G. (1987) Contribution à l'étude de la force articulatoire des occlusives du coréen à partir de méthodes experimentales, *Travaux de l'Institut de Phonétique de Strasbourg, 19,* 37-70

Port R.F., Mitler F.M. (1980) Segmental features and implementation in acquisition of English by Arabic speakers, *Journal of Phonetics, 11,* 219-229

William S L. (1977) The voicing contrast in Spanish, *Journal of Phonetics, 5,* 169-184

QUELQUES ASPECTS DE L'HYPOARTICULATION EN FRANÇAIS SPONTANE

Danielle DUEZ

CNRS URA 261, Laboratoire Parole et Langage, Aix en Provence

Tel.: 42 95 36 32 - Fax : 42 59 50 96 -e-mail: duez@lpl.univ-aix.fr

ABSTRACT

Some of the effects of underarticulation on the acoustic structure of intervocalic voiced stops in spontaneous speech were investigated in both perceptual and spectrographic studies. A certain number of consonants were shown to be reduced or assimilated with adjacent vowels or nearby consonants. These changes may be the result of decreases in the magnitude of individual gestures or the overlapping production of speech gestures. The place of articulation tends to be maintained in reduced or assimilated consonants.

1.INTRODUCTION

Dans son modèle hypo/hyperarticulation, Lindblom (1990) décrit le processus de production de parole naturelle comme un processus où les locuteurs ajustent leurs gestes articulatoires aux besoins communicatifs et situationnels. Dans les situations de communication formelle, où les contraintes de réception dominent, le locuteur tend à hyperarticuler afin d'être intelligible pour tous; en revanche, dans les situations de conversation informelle, où les contraintes de communication sont souvent peu sévères, le locuteur aura tendance à économiser ses gestes de production et hypoarticuler. Cette économie des gestes articulatoires peut se traduire par un appauvrissement sévère de la structure acoustico-phonétique du mot *réalisé* comparée à sa forme *intentionnelle*. Cependant, comme la quantité d'information minimale nécessaire à l'accès lexical varie avec la localisation des segments dans la phrase, dans le mot, et dans la syllabe, les gestes phonétiques ne seront pas simplifiés de manière égale et uniforme.

Certains des effets de l'hypoarticulation sur la structure acoustique des occlusives voisées intervocaliques de la parole spontanée sont examinés ici. Deux hypothèses sont testées. La première est que l'hypoarticulation se traduit par la réduction et l'assimilation au contexte des occlusives voisées. La réduction est définie comme un processus d'affaiblissement qui conduit à une moindre constriction de la consonne: une occlusive est changée en affriquée ou constrictive, ou une fricative devient une sonorante. L'assimilation est un processus qui augmente la similarité entre des segments adjacents (ou plus lointains), le segment assimilé pouvant être remplacé par le segment dominant et assimilant. La seconde hypothèse est que les patrons de réduction et d'assimilation obéissent à certaines règles et exhibent certaines régularités.

La première partie de cette étude consiste en une analyse perceptive d'occlusives intervocaliques extraites de conversation avec les voyelles adjacentes. Les consonnes perçues sont ensuite comparées avec les consonnes prototypiques. Deux groupes sont définis : consonnes perçues comme voulues (1) et consonnes perçues différemment (2). Les consonnes perçues sont ensuite classées selon le voisement, le mode et lieu d'articulation, et en relation avec les voyelles adjacentes, et les consonnes précédente et subséquente. Une analyse spectrographique des consonnes constitue la seconde partie de cette étude : elle est limitée à quelques indices acoustiques pertinents. Ne seront présentés ici que les résultats concernant les consonnes perçues différemment.

2. ANALYSE PERCEPTIVE

2.1. Méthode

2.1.1. Locuteurs.

Deux heures de conversation de deux locuteurs français ont été enregistrées en chambre sourde à l'Institut de Phonétique de Stockholm avec un Revox PR/99 et un microphone de type Brüel et Kjaer. Ces locuteurs âgés respectivement de 20 et 25 ans ne présentaient pas de trouble d' élocution ou auditif connu, ils avaient fait des études universitaires et étaient natifs du Nord de la France et de Paris.

2.1.2. Procédure d'échantillonnage.

Les deux heures de conversation ont été transcrites par deux phonéticiens. Toutes les occlusives voisées intervocaliques ont été localisées dans la transcription et la présence de voyelles adjacentes a été contrôlée auditivement. Des séquences de type V_1CV_2 (ou C est une occlusive voisée, et V_1 et V_2 toute voyelle orale ou nasale) ont été échantillonnées et sauvegardées. Les limites des séquences coïncident avec le milieu de V_1 et le milieu de V_2, déterminé sur tracé oscillographique et spectrogramme à bande large. La durée moyenne d'un stimuli est de 180 ms, la fréquence d'échantillonnage de 16 kHz.

2.1.3. Stimuli.

Les deux séries de stimuli consistent en 361 séquences V1CV2 pour Loc1 (217 /d/, 46 /b/,18 /g/ et 80 autres consonnes utilisées comme distracteurs) et 372 séquences V1CV2 pour Loc2 (208 /d/, 65 /b/, 19 /g/, et 80 distracteurs. L'ordre de présentation des consonnes est aléatoire. Chaque stimulus est répété deux fois. L'intervalle entre les répétitions est d'1 s, et de 3 s entre deux stimuli. La durée totale de chaque série est de 40 minutes.

2.1.4. Procédure d'écoute.

Vingt cinq auditeurs sans aucun problème d'audition connu ont participé à l'expérience. Au cours de deux sessions, ils ont été testés individuellement en chambre sourde. Leur tâche consistait à écrire tout ce qu'ils entendaient: consonne, groupe de consonnes, pas de consonne, ainsi que les voyelles adjacentes. Il leur a été précisé que les stimuli étaient extraits de conversations.

2.2. Analyse

A chacun des stimuli est assignée une identification correspondant à la consonne reportée par plus de 50% des locuteurs (au moins 13). Dans le premier groupe, il y a accord entre la consonne prototypique et la consonne perçue par 50% de auditeurs. Dans le second groupe, quatre cas sont considérés : (1) une consonne différente est perçue par au moins 13 auditeurs (ex /d/ perçu comme /n/, (2) la majorité identifie un lieu d'articulation (ex dental), 3) la majorité n'identifie aucune consonne ou aucun lieu d'articulation, 4) la consonne est reportée comme omise. L'identification des consonnes est examinée en fonction de la nature des voyelles adjacentes, du voisement, mode et lieu d'articulation des consonnes C1 et C2 dans les séquences C1VC1VC2

2.3. Résultats

Pour un certain nombre de stimuli, la consonne perçue est différente de la consonne prototypique. La répartition est la suivante : Loc1 (/b/: 15/65; /d/: 38/208; /g/: 1/19); Loc2 (/b/: 8/46; /d/: 28/217; /g/: 1/18).

2.3.1. Consonnes identifiées comme nasales

Un certain nombre de /b/ et /d/ sont identifiés comme /m/ et /n/ : 2 /m/ (Loc1) et 1 /m/; (Loc2) ; 10 /n/ (Loc1) et 12 /n/ (Loc2). La majorité des /m/ et /n/ sont localisées dans un contexte vocalique nasal à l'exception de deux /d/ perçus comme /n/ (Loc1). Notons cependant que l'un de ces /d/ est précédé par un /m/. Le contexte vocalique a un effet significatif sur l'identification des /d/ comme /n/ pour les deux locuteurs, et pour l'identification des /b/ pour Loc1. (/d/ Loc1: X^2=13.3; Loc2: X^2=79, p=0.0001; /b/, Loc1: X^2=13, p=0.001, Loc2: n.s.). L'identification des occlusives nasales reflète probablement les influences assimilatrices des voyelles nasales adjacentes ou d'une consonne nasale plus lointaine.

2.3.2. Consonnes identifiées comme latérales

Un certain nombre de /d/ localisés dans un contexte de voyelles antérieures sont identifiés comme /l/: 5 (Loc1) et 1 (Loc2), Comme ces /l/ sont hors de tout contexte de consonnes latérales, l'identification des auditeurs ne peut être vue comme le résultats d'une assimilation, mais plutôt comme le résultat d'un processus de réduction de l'occlusive en latérale.

2.3.4. Consonnes identifiées comme occlusives non voisées

Neuf /d/ sont identifiés comme /t/ : 2 (Loc1) et 7 (Loc2). Sept de ces /t/ sont localisés dans un contexte consonantique non voisé. Le faible nombre de /t/ ne permet cependant aucune généralisation sur l'assimilation des consonnes au contexte. Les deux /t/ localisés dans un contexte de consonnes voisées peuvent être le résultat d'erreurs de perception dues à la durée à la procédure d'excision adoptée.

2.3.5. Consonnes perçues comme fricatives

Cinq /d/ sont perçus comme /z/ (1: Loc1 et 3: Loc2) et /s/ (1: Loc1); 14 /b/ sont perçus comme /v/ (Loc1: 9; Loc2: 5)/. Le faible nombre de cas ne permet pas de relier cette

identification à un contexte donné, notons que les /z/ et /s/ tendent à être dans un contexte fricatif. Un /g/ est aussi identifié comme /v/: il y a alors perte du lieu et mode d'articulation.

2.3.6. Autres consonnes

Pour cinq /b/, seul le lieu d'articulation labial est identifié (Loc1: 4 et Loc2: 1). Les auditeurs identifient le lieu d'articulation mais ne peuvent décider si la consonne est /b/, /v/ ou /w/, en partie à cause d'une information insuffisante. Un /b/ est identifié comme /w/. Il peut être vu comme un affaiblissement du /b/ en semi-voyelle correspondante, il peut être aussi une erreur d'identification due à une durée vocalique contextuelle incorrecte. Le 13 /d/ pour lesquels seul le lieu d'articulation dental est identifié sont dans un contexte de voyelles antérieures. Le plus souvent, les auditeurs n'ont pu décider si ces consonnes sont des /d/, /l/, /z/, ou /n/. Elles peuvent avoir été changées en "approximantes", Pour 7 /d/ (Loc1: 1; Loc2 :1), il n'y aucune identification de consonne, ni d'un lieu d'articulation. Trois /d/ et un /g/ (Loc1) sont omis: il peut s'agir d'un cas extrême de réduction.

3. ANALYSE ACOUSTIQUE

3.1. Méthode

Des spectrogrammes à bande large sont générés pour chacun des stimuli. Les consonnes prototypiques étant des occlusives voisées, la présence d'une barre de voisement, et d'une explosion est contrôlée pour chacun des stimuli. Dans le cas positif, la durée de l'occlusion et du VOT (intervalle compris entre la détente de l'explosion et l'attaque de la voyelle) sont notés.

La présence de formants de fréquence moyenne est contrôlée pour chacune des occlusives localisées dans un contexte de voyelle nasale, leurs limites sont placées au point de changement spectral maximal, leur fréquence et leur durée sont mesurées. Trois degrés de nasalisation sont définis : pas de nasalisation (pas de formants de fréquence moyenne), nasalisation partielle (présence d'une occlusive séparée), nasalisation totale (pas d'interruption des formants de fréquence moyenne). Les latérales sont caractérisées par des formants de fréquence moyenne plus forts que pour les nasales, la présence de tels formants est vérifiée pour les latérales et les consonnes identifiées comme /d/ dans un contexte de voyelles orales. L'on sait que le VOT est l'une des caractéristiques des occlusives non voisées de la parole lue. Le VOT des /t/ est comparé avec celui des /d/. Le taux de passages par zéro est calculé sur le signal original et le signal préaccentué pour les fricatives et les /b et d/. La présence de formants est également contrôlée : les fricatives peuvent également être caractérisées par des formants, le spectre des /v et z/ est comparé avec celui des /b et v/. Le taux de passages par zéro est aussi calculé pour les /b et d/ dont seul le lieu d'articulation est identifié, la présence de formants visibles est notée. Dans les cas où aucune consonne ni aucun lieu d'articulation ne sont identifiés, l'hypothèse que la consonne a été omise est testée.

3.2. Résultats

3.2.1. Consonnes identifiées comme nasales

Un chevauchement complet des formants de fréquence moyenne et de l'occlusion caractérise les consonnes perçues comme /m et n/ puisqu'un seul /n/ n'a pas de formants visibles. Un tel chevauchement est rare pour les /b/ (un seul cas) et les /d/ (4 cas).

3.2.2. Consonnes identifiées comme latérales.

Toutes les consonnes identifiées comme /l/ exhibent des formants vocaliques, contrairement à la majorité des consonnes étiquetées comme /d/. Les /l/ ont une durée brève (durée moyenne : 37 ms).

3.2.3. Consonnes identifiées comme occlusives non voisées

Il y a des vibrations laryngiennes continues pour les consonnes identifiées comme /t/. Les VOT sont groupés autour de 0-15 ms pour les /d/ et autour de 30 ms pour les /t/. Une durée intermédiaire de 25-30 ms est obtenue pour deux /t/ et quelques /d/ : elle peut représenter un seuil de perception entre les /t/ et les /d/ en français spontané.

3.2.4. Consonnes identifiées comme fricatives.

Seuls les résultats obtenus pour le taux de passages par zéro sur le signal préaccentué sont reportés : ceux obtenus sur le signal original ne sont pas significatifs. Un taux inférieur à 0.2 caractérise la plupart des /b/ et /v/. Deux /v/ ont cependant un taux supérieur à 0.4. Un taux élevé est noté également pour un /s/ , un /z/ et 14 /d/ alors que deux /z/ ont un taux inférieur à 0.2. Le taux de passages par

zéro n'est probablement pas un paramètre pertinent pour comparer les occlusives et les fricatives. Dans la parole spontanée, les occlusives tendent à être partiellement fricatisées, en particulier devant une voyelle antérieure. En revanche, des formants sont visibles pour 10 /v/ et pour les /z/,suggérant que ces fricatives présentent des caractéristiques d'approximantes. Pour le /g/ identifié comme /v/ on ne relève aucun formant visible et le taux de passages par zéro est de 0.2, ce qui est en accord avec l'idée que le /v/ est une confusion.

3.2.5. Autres consonnes

Paradoxalement, on ne relève aucun formant visible pour les cinq labiales. Toutes présentent une occlusion voisée, mais pas d'explosion. Cette configuration est en accord avec l'hypothèse que ces labiales sont le résultat de confusions induites pas une durée vocalique contextuelle incorrecte. En revanche, les /d/ identifiés comme dentales exhibent tous des formants de fréquence moyenne, ils sont aussi très brefs (durée moyenne: 34 ms). Quand aucune consonne n'est reportée, il y a continuité des formants vocaliques, ce qui suggère que la consonne a été omise. On observe la même chose pour les consonnes non identifiées.

4. CONCLUSIONS

Environ 20 % des plosives voisées intervocaliques présentées dans les séquences VCV n'ont pas été identifiées. Les erreurs de production fréquentes dans la parole spontanée peuvent en être l'une des causes. La technique d'excision adoptée peut être une autre cause: les séquences VCV ont été extraites de la parole spontanée et la durée des voyelles adjacentes a été tronquée. L'information apportée par la durée des voyelles adjacentes peut avoir manqué dans le cas où la perception des contrastes consonantiques peut se fonder sur la durée de la voyelle précédente et/ou de la voyelle subséquente.

Les résultats obtenus suggèrent cependant que la perte d'information est majoritairement le résultat d'un chevauchement plus marqué des gestes de parole (Browman et Goldstein, 1990 et 1992). La nasalisation des occlusives est sans doute le résultat du chevauchement de l'abaissement du vélum et du geste de fermeture. Au niveau acoustique, il n'y a pas d'interruption des formants de fréquence moyenne; au niveau perceptif, il y a identification d'un /m/ ou d'un /n/.

Les données acoustiques et perceptives montrent également des cas d'affaiblissement d'occlusives en fricatives ou en approximantes. Ces changements peuvent être dus à une réduction de l'amplitude des gestes articulatoires, réduction qui conduit à une fermeture incomplète (Browman et Goldstein, 1990 et 1992). Par exemple, le changement du /b/ en /v/ ou /w/ reflète une réduction dans l'amplitude du geste de fermeture, plus marquée dans le cas du /v/. Le changement du /d/en /l/ ou en semi-voyelle reflète également une réduction dans l'amplitude du geste de fermeture. L'effacement du /b/ ou du /d/ peut être le résultat de l'effacement du geste de fermeture.

Le lieu d'articulation tend à être maintenu dans les consonnes assimilées ou réduites. Les gestes consonantiques peuvent être définis comme des gestes structurés, transversaux, produits à des lieux spécifiques et permettant une fermeture complète (Carré et Mrayati, 1990), Dans la parole spontanée, les gestes consonantiques peuvent être incomplets ou mal réalisés, cependant, comme ils sont transversaux, ils gardent leur lieu d'articulation.

5. BIBLIOGRAPHIE

Browman, C.P. et Goldstein, L (1990). Tiers in articulatory phonology, with some implications for casual speech. In *Papers in Laboratory Phonology I: between the grammar and physics of speech,* (M. Beckman, editor), pp 341-376. Cambridge, G.B: The Cambridge University Press.

Browman, C.P. et Goldstein, L. (1992) Articulatory phonology: An overview, *Phonetica*, n°49, 155-180.

Carré, R. et Mrayati, M. (1990) Articulatory-acoustic-phonetic relations and modelling, regions and modes. In *Speech production and speech modelling* (W.J. Hardcastle and A. Marchal, editors), Vol 55, pp. 211-240. NATO ASI Series. Dordrecht, Boston and London: Kluwer Academic Publishers.

Lindblom, B. (1990) Explaining phonetic variation: a sketch of the H and H theory. In *Speech production and speech modelling* (W. Hardcastle and A. Marchal, editors), Vol 55, pp. 403-439. NATO ASI Series. Dordrecht, Boston and London: Kluwer Academic Publishers.

STRATEGIES D'HESITATION PROPRES AUX LOCUTEURS DANS LE FRANÇAIS SPONTANE MEDIATISE

Zsuzsanna FAGYAL

Institut de Phonétique de l'Université de Paris III
University of Pennsylvania Linguistics Department
fagyal@unagi.cis.upenn.edu

Résumé

Cette étude montre que les locuteurs appliquent différentes 'stratégies' d'hésitation pour maintenir la fluidité de leur parole énoncée de mémoire. L'analyse de quatre types de pause sonore (allongement vocalique, pause remplie, répétition et faux-départ) d'environ quatre cents énoncés issus d'entretiens télévisés et radiophoniques indiquent que même à des vitesses de locution comparables, les locuteurs étudiés privilégient certaines combinaisons simples et multiples de pauses sonores plutôt que d'autres. Les différentes stratégies d'hésitation sont discutées en termes d'audio-contrôle, de débit de parole et de contexte situationnel.

Mots-clés : *hésitation, pause remplie, allongement, répétition, faux-départ*

INTRODUCTION

Depuis les travaux de Goldman-Eisler [1] concernant l'importance cognitive des pauses d'hésitation dans la parole, de nombreuses études se sont consacrées à leur analyse. Levelt [2] décrit le phénomène en terme d'audio-contrôle *(audio-monitoring)* et il montre que le locuteur, encodant et produisant la parole de façon simultanée, exerce un contrôle—et si nécessaire effectue des corrections—sur son discours. Malgré l'attention considérable portée à la classification, aux fréquences d'occurrence et à la distribution syntaxique de ces pauses, peu d'études examinent dans quelle mesure leur choix dépend du locuteur.

Les aspects individuels des phénomènes d'hésitation sont abordés par Maclay et Osgood [3]. Une forte corrélation est établie entre le débit et l'occurrence de certaines pauses sonores, mais *la préférence relative* (p.36) des locuteurs pour l'un ou l'autre type de pause n'est pas interprétée de manière systématique. Grosjean et Deschamps [4] notent que *chaque sujet utilise un langage hésitant lors des descriptions, en faisant appel à une combinaison des différentes variables simples qui lui est propre* (p.217), mais cette observation est dérivée indirectement de l'étude focalisant sur les relations entre hésitations et situation *(type d'encodage).* Duez [5] observe également des tendances individuelles chez des locuteurs prononçant différents types de discours politique, mais elle ne trouve pas le facteur individuel prédominant. Les stratégies d'hésitation individuelles sont encore moins étudiées du point de vue de la complexité ou du 'regroupement' des pauses d'hésitation. Le phénomène est connu, mais il constitue rarement un objet d'étude.

La présente analyse suggère que les locuteurs appliquent des stratégies d'hésitation différentes à des débits comparables et dans des situations similaires. L'hésitation serait donc également fonction du locuteur.

2. CORPUS

Le corpus est constitué de 361 énoncés relevés dans 41 minutes de parole d'entretien médiatisé, enregistrée entre 1975 et 1984 avec quatre écrivains : R. Barthes (B), M. Duras (D), F. Giroux (G) et M. Yourcenar (Y). Les entretiens télévisés (D et Y) et radiophoniques (B et G) représentent une situation de parole spontanée 'formelle'. Les extraits ont préalablement servi de corpus d'étude dans deux thèses [6] [7][1].

3. METHODE ET DEFINITIONS

Les prises de parole ayant la longueur minimale d'un énoncé ont été segmentées en **pauses silencieuses** (PSils) et en séquences sonores (SSons). Le seuil de coupure des PSils a été calculé pour chaque extrait à la base de la longueur moyenne de 50 arrêts d'occlusive intervocalique (voir [5]). Dans les quatre extraits, les durées des seuils sont comprises entre 140 ms et 250 ms. La perception des PSils détectées selon cette méthode acoustique a été confirmée par deux auditeurs phonéticiens.

La **pause remplie** (PR) (∂) se réalise en français par l'allongement excessif de la voyelle neutre [∂] et la décroissance progressive de la fréquence fondamentale avec ou sans perturbations *(creaky voice)*. Sa durée varie en fonction du type de segment adjacent (présence ou absence de PSil), mais elle peut être 5 à 10 fois supérieure à la durée des syllabes accentuées [8].

" .. il appartenait à cette bourgeoisie ∂ qui n'a pas de problèmes d'argent. (G)"

La **répétition** (RT) (_) simple ou multiple d'un segment de longueur variable est considérée comme un phénomène d'hésitation lorsque le segment repris n'apporte aucune signification nouvelle à l'énoncé [3]. Ce critère permet de séparer cette pause sonore des autres phénomènes de répétition, comme les reprises d'adverbes du type «très très content», à base de critères intonatifs [9]. Ainsi, lorsque les réalisations mélodiques de l'amorce et de la reprise sont différentes, le segment repris n'est pas qualifié de répétition, car il s'agit d'un apport d'information marqué par des moyens intonatifs [10].

" ... je n' suis pas vraiment le le seul fondateur (B)"

L'allongement (AL) (:) est un moyen d'hésitation privilégié en français qui a une forte tendance à l'allongement final [12]. L'hésitation par allongement, accroissement excessif de la durée syllabique, se produit principalement en fin de constituants intonatifs majeurs [4] [5] [6] [8]. Bien que la durée des consonnes puisse être également allongée [6], seul l'allongement vocalique est étudié ici :

" .. que : un livre est toujours pris un peu comme une certaine culpabilité (D)"

Les **faux-départs** (FD) (/) sont des 'accidents de production' *(slips of the tongue)* de la parole non lue. Ils se distinguent des manifestations pathologiques, comme le bégaiement, dans la mesure où ils 'respectent' les frontières syllabiques [11]. Les faux-départs peuvent être grammaticaux ou lexicaux, repris ou non repris. Lorsqu'ils sont repris, ils apportent une correction, donc une nouvelle information à l'énoncé (voir répétitions).

" .. j'avais fait / j'avais suivi quelques cours d'archéologie... (Y)"

4. STRATEGIES INDIVIDUELLES

Les pourcentages d'occurrence des pauses sonores (fig.1) indiquent deux stratégies principales d'hésitation chez les locuteurs : l'allongement ou l'insertion de pauses remplies. Les stratégies des locuteurs D et B sont opposées en ce qui concerne ces deux PSons. Les ALs semblent être privilégiés par D (40%), mais ils sont d'une faible proportion dans la parole de B (16%). D n'a recours aux PRs que dans 8,7% des cas, alors que celles-ci représentent plus que la moitié (50,8%) des occurrences totales des pauses sonores dans la parole de B.

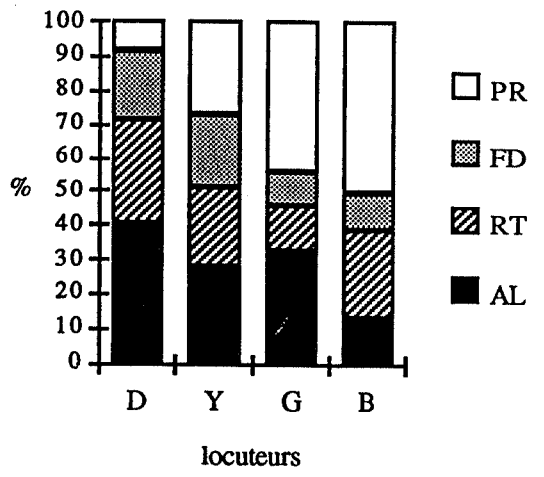


Figure 1. Pourcentages des pauses sonores dans la parole de quatre locuteurs. Pauses remplies (PR), faux-départs (FD), répétitions (RT) et allongements (AL).

La différence entre les locuteurs D et B à l'égard des ALs peut s'expliquer par leurs débits moyens (tabl.1) : D parle en moyenne moins vite (130 mots/mn) que B (167 mots/mn) pendant l'entretien. D aurait donc tendance à ralentir plutôt qu'à répéter ou à insérer des segments dans sa parole en 'hésitant' [3] [4]. En revanche, le débit lent de D n'explique pas la présence d'un grand nombre de RTs, phénomène contraire à certaines observations faite pour l'anglais [3]. Il paraît encore plus 'surprenant' que D, parlant pourtant à un débit relativement lent, ait plus de faux-départs dans son discours que B (20% et 10,4%).

En laissant momentanément de côté la question des faux-départs, l'ordre décroissant de 'préférence relative' des PSons par les locuteurs D et B peut être établie comme suit :

D -> **AL**>*RT*>PR
B -> **PR**>*RT*>AL

Dans la parole de D, le nombre d'ALs est significativement plus élevé que celui des autres pauses sonores. On observe le même phénomène dans la parole de B quant au

nombre de PRs (tabl.1). Les RTs occupent la seconde place dans l'ordre de préférence des deux locuteurs.

La locutrice G applique la combinaison des deux stratégies précédentes avec une légère dominance des PRs. Ainsi, parlant à un débit intermédiaire (148 mots/min), elle recourt souvent à l'allongement (35% des cas), mais elle tend plutôt à insérer des pauses remplies (48%) :

G -› **PR**›*AL*›RT

Sa stratégie diffère significativement de celles de D et de B (tabl.1). Son discours paraît également plus 'optimal' avec un bon compromis entre débit relativement rapide, qualité (peu de RTs) et fluidité (peu de FDs) du discours.

	Types de PSon					N total	Débit
	PR	FD	RT	AL	p*	PSons	mot/mn
D	10	23	36	46	0,001	115	130
Y	25	20	21	26	*0,76*	92	149
G	37	8	11	27	0,001	83	148
B	63	13	32	16	0,001	124	167
p*	0,01	0,002	*0,06*	0,002			

* (tests chi^2) significatif à p‹0,005

Tableau 1. Nombres des PSons et débits de locution dans la parole de quatre locuteurs.

La locutrice Y ne semble opter pour aucune stratégie en particulier ; elle combine les trois précédentes. La différence entre le nombre relatif des quatre pauses sonores n'est pas significative dans son cas (tabl.1). Au débit comparable à la parole de G, celle de Y montre une distribution équitable entre les trois types de Psons étudiées. Cette 'absence' de stratégie d'hésitation est propre à la locutrice Y :

Y -› AL PR RT

Ces règles de combinaison se retrouvent également dans les *regroupements* [4] plus ou moins longs des pauses sonores *que le locuteur n'hésite pas à accumuler lorsqu'il a besoin d'un temps assez élevé pour programmer l'unité suivante* [5] (p.74). L'exemple suivant représente la suite de pauses la plus longue relevée dans le corpus :

"... pas trop ∂ # ∂ j' ∂ j'essaie de j'essaie de de n'pas trop m'en occuper.." (B)

PR+PSil+PR+RT(PR)+RT+RT

Cette longue suite de pauses sonores a été prononcée par le locuteur B qui, conformément à sa stratégie, privilégie les PRs (∂) et les RTs (_) comme moyen d'hésitation. Les PSils (#) pouvant servir d'arrêts d'hésitation sont également comptabilisées lors de l'analyse des regroupements, bien que leur étude ne puisse être abordée dans ces pages.

Selon notre hypothèse, lorsqu'un locuteur est amené à accumuler plusieurs pauses sonores pour prolonger le délais de l'encodage tout en gardant la parole, il fait appel aux pauses sonores privilégiées par sa stratégie individuelle. Afin de vérifier cette hypothèse, les trois pauses sonores ont été analysées en fonction de leurs occurrences simples et multiples avec des PSils et d'autres pauses sonores. Trois catégories d'occurrences ont été retenues : simples (=1), double (=2) et multiple (›2). L'hypothèse des stratégies individuelles semble être confirmée chez l'ensemble des locuteurs (tabl.2). D, dont la stratégie d'hésitation est basée principalement (en gras) sur les ALs et dans une moindre mesure (en italique) sur les RTs, combine ces deux pauses pour produire des suites doubles (22 ALs mais 2 PRs) et triples (15 ALs, 21 RTs mais 3 PRs). Les valeurs sont distribuées de manière égale entre les trois catégories et les trois types de pauses pour la locutrice Y, n'ayant pas de stratégie d'hésitation particulière. On trouve également de nombreuses suites doubles chez les locuteurs G et B qui font appel au paramètre principal (en gras) de leur stratégie d'hésitation : la pause remplie. Conformément à ses 'préférences', G a moins souvent recours aux RTs, alors que B fait moins appel aux ALs (en italique). On trouve de nombreuses occurrences simples (=1) de PRs chez D comme chez B. Ceci s'explique probablement par la nature de la PR qui, ne possédant aucun sémantisme intrinsèque, se combine facilement avec tout élément de la chaîne lexicale.

DISCUSSION ET CONCLUSION

La question des faux-départs, laissés de côté lors de l'analyse des différentes stratégies, reste à discuter en premier lieu. Contrairement aux trois pauses d'hésitation étudiées auparavant, le faux-départ ne peut pas être **'choisi'** comme 'moyen stratégique' d'hésitation. Il est, au contraire, l'indice de **l'audio-contrôle**, c'est-à-dire de la capacité (possibilité ou volonté) du locuteur de contrôler (reprendre et corriger) son discours. Dans des situations d'entretien à deux, comme celles dans ce corpus, le locuteur n'a pas de

	Simples (=1)				Doubles (=2)				Multiples (>2)				N
	AL	RT	PR	FD	AL	RT	PR	FD	AL	RT	PR	FD	total
D	**9**	*10*	5	8	**22**	*5*	2	6	**15**	*21*	3	9	115
Y	7	7	5	12	6	5	11	-	13	9	9	8	92
G	*8*	5	**11**	2	*9*	1	**24**	3	*10*	5	**2**	3	83
B	5	*19*	**18**	8	9	*4*	**36**	1	2	*9*	**9**	4	124
N total	29	34	39	30	46	15	73	10	40	44	23	24	

Tableau 2. Regroupements de pauses sonores dans la parole de quatre locuteurs. Les chiffres renvoyant aux types de pauses 'privilégiés' par le locuteur sont marqués en gras et en italique.

raison particulière de surveiller de près la fluidité de son discours, n'étant pas menacé d'interruptions. Ceci peut expliquer pourquoi on observe à la fois beaucoup de faux-départs à un débit relativement lent (D) et peu de faux-départs à un débit plus élevé (B).

L'analyse des occurrences de pauses sonores relevées dans la même situation d'énonciation auprès de quatre locuteurs indique que le choix des paramètres d'hésitation est fonction du locuteur. Ceux-ci appliquent des combinaisons de pauses qui leur sont propres, lorsqu'ils sont amenés à prolonger le temps d'encodage en gardant la parole. Nos résultats soutiennent l'idée de Heike [13] qui considère que chaque locuteur doit trouver son propre compromis entre le délais d'encodage nécessaire et le maintien de la fluidité du discours. Cette dernière contrainte devient impérative, par exemple, en situation de débat. Dans une telle situation, le locuteur—bien que 'préférant' l'hésitation par allongement—peut être amené à modifier sa stratégie: appliquer plus de répétitions et 'risquer' plus de faux-départs [6]. Il nous semble que l'étude des hésitations dans la parole doit tenir en compte des stratégies individuelles des locuteurs avec et sans l'influence du contexte situationnel.

(1) Je remercie Piet Mertens d'avoir mis à ma disposition les enregistrements avec les locuteurs B et G.

BIBLIOGRAPHIE

[1] GOLDMAN-EISLER, F. (1968), Psycholinguistics : Experiments in Spontaneous Speech. Academic Press, London and New York.

[2] LEVELT, W. J. M. (1983), Monitoring and self-repair in speech. *Cognition*, 14, 41-104.

[3] MACLAY, H. - OSGOOD, E. Ch. (1959), Hesitation Phenomena in Spontaneous Speech. *WORD* (15) 1, April, 19-44.

[4] GROSJEAN, F. - DESCHAMPS, A. (1973) Analyse des variables temporelles du français spontané II. *Phonetica* 28 : 191-226

[5] DUEZ, D. (1991), La pause dans la parole de l'homme politique. Editions du CNRS, Aix-Marseille.

[6] FAGYAL, Zs. (1995), Aspects phonostylistiques de la parole médiatisée lue et spontanée. Thèse de doctorat, Université de Paris III Sorbonne Nouvelle.

[7] MERTENS, P. (1987), L'intonation du français : de la description linguistique à la reconnaissance automatique. Thèse de doctorat, Université Catholique de Leuven.

[8] GUAÏTELLA, I. (1992), Analyse acoustique, perceptive et fonctionnelle des hésitations vocales en parole spontanée. *Actes des XIXes JEP*, Bruxelles, 171-176.

[9] O'SHAUGHNESSY, D. (1992), Analysis of false starts in spontaneous speech. *Proceedings of the ICSLP*, Banff (Alberta), 931-934.

[10] BROWN, G. (1983) Prosodic Structure and the Given/New Distinction. in : A. Cutler - D. Ladd (eds.), Prosody : Models and Measurements. Springer Verlag, 67-77.

[11] ZELLNER-BECHEL, B. (1992), Le be-bégayage et euh..., l'hésitation en français spontané. *XIXe JEP*, Bruxelles, 171-176.

[12] VAISSIERE, J. (1991), Rhythm, accentuation and final lengthening in French. in : J. Sundberg et al. (eds.), Music, Language, Speech and Brain. Macmillan Academic Press, London 108-120.

[13] HIEKE, A. (1981) A Content-processing view of hesitation phenomena. *Language and Speech*, 24, 2, 147-160.

UN MODÈLE TRIDIMENSIONNEL POUR L'ÉTUDE DE LA VARIATION ET DES CHANGEMENTS PHONÉTIQUES EN COURS

Isabelle MALDEREZ

Université Paris 7 - Denis Diderot, UFR Linguistique, Laboratoire de Phonétique, EA 333

isabelle.malderez@linguist.jussieu.fr

Abstract

I intend to present a tri-dimensional model for research on phonetic variation and its application to the study of dynamic synchrony of rounded oral vowels in French of Ile-de-France. This approach conjointly examines the three domains of appearance of ongoing sound changes that are production, perception and spelling. Production's study is very well defined in variationist linguistics: I will explain here especially the two other matchings of my approach. Janson [6] showed that category perception can reveal the ongoing changes between phonems by comparing two populations which represent successive generations. The experiment consists in a test where subjects have to categorize a set of synthetic stimuli stemed from a linear interpolation of the two targets' parameters. As for the spelling, Fónagy [4] considers that spelling mistakes of young children often reflect their "phonological conception". So, systematic collect of atypical lexical spelling mistakes would reveal the emerging sound changes or "below the level of social awareness" [7].

LES VARIABLES DEPENDANTES

l'antériorisation de /O/

Vaugelas [15] donne des exemples de mots qui étaient prononcés comme s'ils étaient écrits avec un e. Commencer devient quemencer et pour lors devient pour l'heure. Nous pouvons trouver d'autres exemples chez Martinet [13] ou Walter [16].

la postériorisation de /Ø/

Walter [16] et Lefebvre, Goudailler et Peretz in Houdebine [5] ont montré que certaines personnes et plus particulièrement les enfants prononcent les /Ø/ postériorisés surtout après /r/.

ANALYSE DE LA PRODUCTION ÉCRITE

Les bases théoriques de cette étude ainsi que la procédure utilisée et les principaux résultats ont été présentés dans Malderez 1994 et renvoient à la 'conception phonologique' de jeunes enfants et d'adultes [4] à qui j'ai proposé une série d'exercices à trous présentant principalement les variables dépendantes dans le contexte favorisant la postériorisation. Les tests ont été effectués dans trois écoles primaires rurales du sud de l'Oise, par 230 enfants.

Il apparait dans cette enquête qu'à l'écrit, en syllabe initiale, les voyelles e et o sont confondues par les jeunes enfants (7-9 ans), surtout dans le contexte /r/_. Cette confusion, pour les plus jeunes enfants, est plus importante que celle qui concerne la paire ou, u. Chez les enfants plus âgés, le pourcentage d'erreur est minime et le contexte n'est pas un critère opératoire pour le test portant sur e et o.

J'ai aussi pu mettre en évidence l'influence d'une prononciation postériorisée dans l'écrit d'adultes en particulier pour le lexème *reblochon*, nom provenant d'un type de fromage dont l'orthographe n'est pas assimilée par tous, qui est souvent prononcé [robloʃɔ̃].

ANALYSE DE PERCEPTION CATÉGORIELLE

Objectif

Il s'agit, grâce à une série de courts tests, de mettre en évidence d'éventuels changements phonétiques en cours, de montrer leur direction et les tranches d'âges touchées par le phénomène [11].

Janson [6] a montré que le test de perception catégorielle permet de mettre en lumière les changements en cours entre deux phonèmes en comparant deux populations représentant deux générations successives. En effet, selon lui, si un changement est en cours, les deux populations n'ont pas les mêmes résultats quant à la catégorisation des stimuli. En particulier, la frontière phonématique dans le continuum reliant les deux phonotypes n'est pas la même pour les deux groupes d'âges. L'expérience consiste donc à faire catégoriser par les sujets une série de stimuli synthétiques issus d'une interpolation linéaire des paramètres des deux phonèmes cibles.

Méthodologie de la synthèse des stimuli

La synthèse des stimuli a été réalisée, avec l'aide de Gérard Bailly à l'Institut de la Communication Parlée. Pour chaque test, nous sommes partis de deux voyelles types. Nous avons choisi pour leur synthèse les 24 paramètres définis dans COMPOST™ [1]. Un programme d'interpolation a ensuite été construit afin de produire 19 stimuli intermédiaires. Il s'agit d'une interpolation linéaire des 24 paramètres des deux voyelles phonotypes. Les 21 stimuli obtenus ont ensuite été triplés. Les 63 stimuli ont été mélangés aléatoirement.

Méthodologie du test de perception

a) Support : Le test se présente donc, pour chaque paire, sous la forme d'une suite de 63 stimuli séparés par 3 secondes de silence. Pour chaque paire, le test dure environ 3 minutes 15 secondes soit un test global de 16 minutes. Un signal musical est joué tous les 9 stimuli, ceci pour faciliter le déroulement du test. J'ai en effet choisi de ne pas faire entendre un numéro avant chaque stimulus : celui-ci aurait pu influencer le choix de l'auditeur. Les feuilles de réponses distribuées aux sujets présentent donc les réponses par blocs de 9.

b) Consignes : Le sujet doit impérativement, pour chacun des stimuli, opter pour une des deux solutions proposées en l'entourant. Je le sensibilise au fait qu'il n'entendra pas forcément autant de fois les deux types de voyelles. Il n'a pas le droit de revenir sur son choix après écoute des stimuli suivants.

c) Sujets : Les 26 sujets ayant passé ce test sont sept "enfants" de moins de 15 ans, huit "jeunes" de 16 à 25 ans, cinq "jeunes adultes" de 26 à 35 ans et six "adultes" de 36 à 45 ans. Cette population est socialement assez homogène : tous les sujets appartiennent aux classes moyennes et sont domiciliés dans l'Oise. Toutes les personnes de la génération "enfants" sont nées dans l'Oise et/ou y résident depuis leur plus tendre enfance. Parmi les autres, certains sont originaires des départements limitrophes ou encore d'autres régions du Nord de France.

Quantification : indice d'antériorité (A)

A chaque stimulus est attribué la valeur (N), nombre de fois où il est perçu comme étant la voyelle antérieure de la paire.

Résultats pour les paires /œ-ɔ/ et /ø-o/

Le facteur "sexe"

La théorie variationiste du changement phonétique prédit une différence de traitement des variables concernées selon les sexes [7, 8, 9]. Cependant, dans le cadre de cette expérience de perception catégorielle, ce facteur n'est pas significatif. Les variables dépendantes ne sont pas liées à la variable indépendante "sexe du sujet", ceci ni dans la population prise globalement ni dans les générations enfants et parents prises isolément. Autrement dit, aucune des oppositions vocaliques étudiées ici ne présente de traitement sexuellement différencié.

Le facteur génération

Pour ces deux oppositions, le rôle du facteur génération est statistiquement significatif. L'espace occupé par les catégories /ø/ et /œ/ est plus important chez la génération parent que chez la génération enfant. Donc, cette étude met en évidence une différence dans le découpage catégoriel des continuums [ø_o] et [œ_ɔ] chez deux générations successives en terme de coupure plus postérieure chez les plus âgés

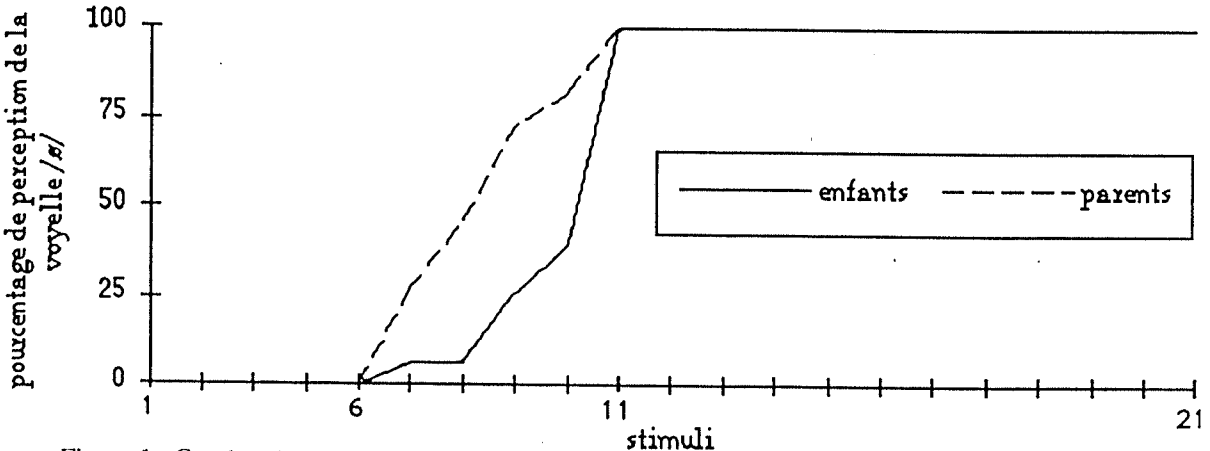


Figure 1 : Courbes de perception catégorielle (indice A') de l'opposition /ø/-/o/ selon les générations

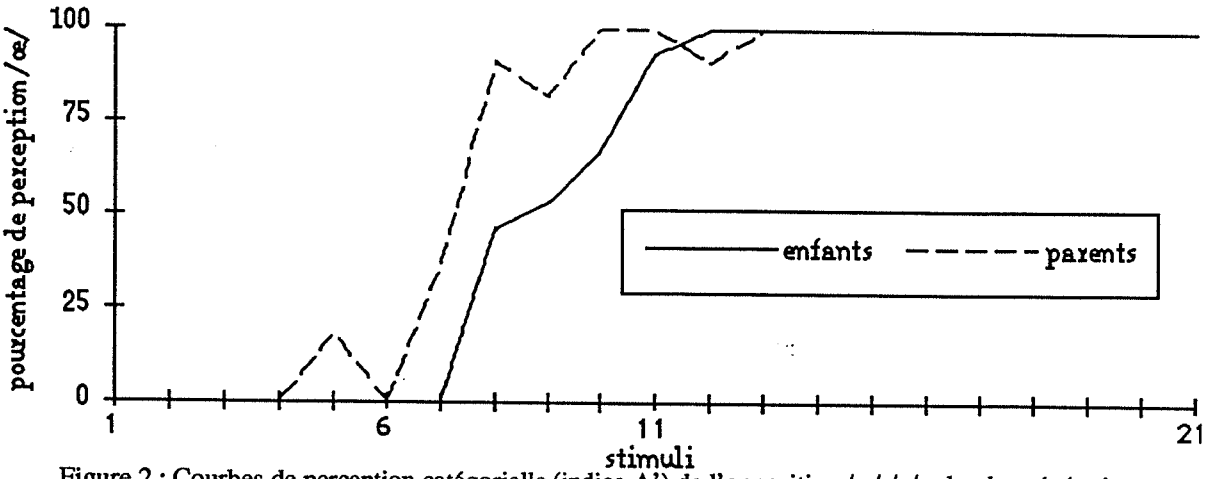


Figure 2 : Courbes de perception catégorielle (indice A') de l'opposition /œ/-/ɔ/ selon les générations

ANALYSE DE LA PRODUCTION

Enregistrements et population

Les enregistrements ont eu lieu au domicile des locuteurs dans une pièce isolée du reste de la famille pour éviter le bruit nuisible aux études acoustiques. Les enregistrements sont constitués de deux parties : d'une part un entretien avec moi-même et d'autre part la lecture de plusieurs textes. Les locuteurs ont lu les textes seuls après les avoir préparés.

L'étude acoustique concerne seulement 8 des 26 personnes ayant passé le test de perception choisis selon deux critères. Premièrement, il s'agit de quatre sujets féminins et de quatre sujets masculins, appartenant aux quatre groupes d'âge. Je me suis arrangée pour que les sujets du même sexe aient une dizaine d'années de différence. J'ai préféré, pour chaque groupe, étudier la voix du locuteur qui présentait les enregistrements de la meilleure qualité sonore.

Méthodologie de l'analyse

J'ai étudié 676 /O/ et 534 /Ø/ ni précédés ni suivis d'une semi-voyelle. Pour la mesure des formants l'analyse cepstrale est la plus efficace de celles que j'ai expérimentées. J'ai utilisé le logiciel FORMANT™ [2].

On trouve dans la littérature de nombreuses propositions de procédures de normalisation des données formantiques. J'ai choisi une méthode simple et éprouvée : j'ai utilisé la procédure de Nearey [14].

Analyse acoustique: Etude des variables indépendantes extra-linguistiques - Voyelle /O/

D'une part, quand on considère l'ensemble des 8 locuteurs, il n'y a pas de production sexuellement différenciée pour cette voyelle. D'autre part, le facteur génération n'est pas opératoire. Enfin, la population n'est pas homogène puisque les productions des différents locuteurs sont statistiquement dissemblables.

Les deux générations présentent des résultats opposés en ce qui concerne les facteurs sexe et locuteurs. Dans la génération parents (325 occurrences), les femmes ne produisent pas de [o] différents de ceux des hommes. Dans la génération enfants (209 occurrences), le facteur sexe est significatif. On peut même dire que les [o] des garçons sont statistiquement plus antérieurs. Enfin, la génération parent est homogène tandis que celle des enfants présente des différences dues aux individus.

Voyelle /Ø/

La population est aussi peu homogène en ce qui concerne la voyelle /Ø/. Le facteur génération est significatif pour cette voyelle. Les plus vieux produisent les voyelles les plus antérieures. En revanche, hommes et femmes produisent le même type de voyelle /Ø/, le facteur sexe n'est pas significatif.

A l'intérieur de chaque génération, les facteurs sexe et locuteur sont significatifs. Tout d'abord, dans la catégorie parents (399 occurrences), ce sont les femmes qui produisent les [ø] les plus antérieurs tandis que, chez les enfants (277 occurrences), se manifeste le résultat inverse. Enfin, les deux générations sont fortement hétérogènes en termes de locuteurs.

CONCLUSIONS

J'ai présenté un modèle de description de la variation phonétique combinant trois approches complémentaires. L'application de cette démarche à l'étude des voyelles moyennes arrondies du français a permis de dégager quelques premiers résultats :

- un plus grand nombre de fautes d'orthographe dans les petites classes que sur une autre paire

réputée stable en français de France;
- une perception différenciée en terme d'âge;
- une production différenciée en terme d'âge pour la voyelle /Ø/.
Cependant, ce travail reste une étude préliminaire de par le nombre de sujets qu'il met en cause et demande à être développé sur une population plus étendue.

BIBLIOGRAPHIE

[1] Bailly, G. & Allissali, M. (1993) Compost : un serveur de synthèse de parole multilingue, *Traitement du signal, 9(4)*, 359-366.
[2] Feng, G. (1983) Détection et mesures numériques de la fréquence fondamentale et des formants du signal de parole, Mémoire de DEA, Institut de Phonétique de Grenoble
[3] Fónagy, Ivan (1964), A Dynamic Approach to Phonetics. Revolution vs. Continuity in the Study of Language, Symposium n°25, Wenner-Gren Foundation for Anthropological Research, Manuscrit.
[4] Fónagy, I. (1989), "Le français change de visage?", *Revue Romane,* 24(2), 225-254.
[5] Houdebine, Anne-Marie (éd.) (1985), *La phonologie de l'enfant français de six ans. Variétés régionales*, Hamburg : Buske.
[6] Janson, Tore (1986), Sound Change in Perception: An Experiment, in *Experimental Phonology*, J.J.Ohala & J.J. Jaeger (eds.), Orlando : Academic Press, 253-260.
[7] Labov, William (1972), *Sociolinguistic patterns*, Philadelphia: University of Pennsylvania Press.
[8]Labov, W. (1992), "La transmission des changements linguistiques", *Langages, 108,* 16-33.
[9] Labov, W. (1994), *Principles of Linguistic Change*, Oxford : Blackwell.
[10] Malderez, Isabelle (1995a), Contribution à la synchronie dynamique du français d'Ile-de-France : le cas des voyelles orales arrondies, Thèse de Doctorat non publiée, Université Paris 7-Denis Diderot.
[11] Malderez, I. (1995b), The Use of a Category-Perception Test in the Study of Ongoing Sound Change, *XIII th ICPhS*, août 1995, Stockolm.
[12] Malderez, I. (1995c), A Tri-dimensional Approach of Study on Sound Change, *New Ways of Analyzing Variation*, october 1995, Philadelpha, University of Pennsylvania. (Manuscrit).
[13] Martinet, André (1958), "C'est joli le Mareuc", *Romance Philology,* 11, pp. 345-355.
[14] Nearey, Terence, (1977), Phonetic feature systems for vowels, University of Connecticut.
[15] Vaugelas, Claude Favre de. (1647), *Remarques sur la langue française utiles à ceux qui veulent bien parler et bien écrire*, Paris : Augustin Courbé & Vve Camusat.
[16] Walter, Henriette (1977), *La phonologie du français*, Paris : P.U.F.
[17] Weinreich, Uriel, W. Labov & M.I. Herzog (1968), Empirical Foundations for a Theory of Language Change, in *Directions for Historical Linguistics*, Austin : University of Texas Press, (3ème ed., 1975. pp. 97-195).

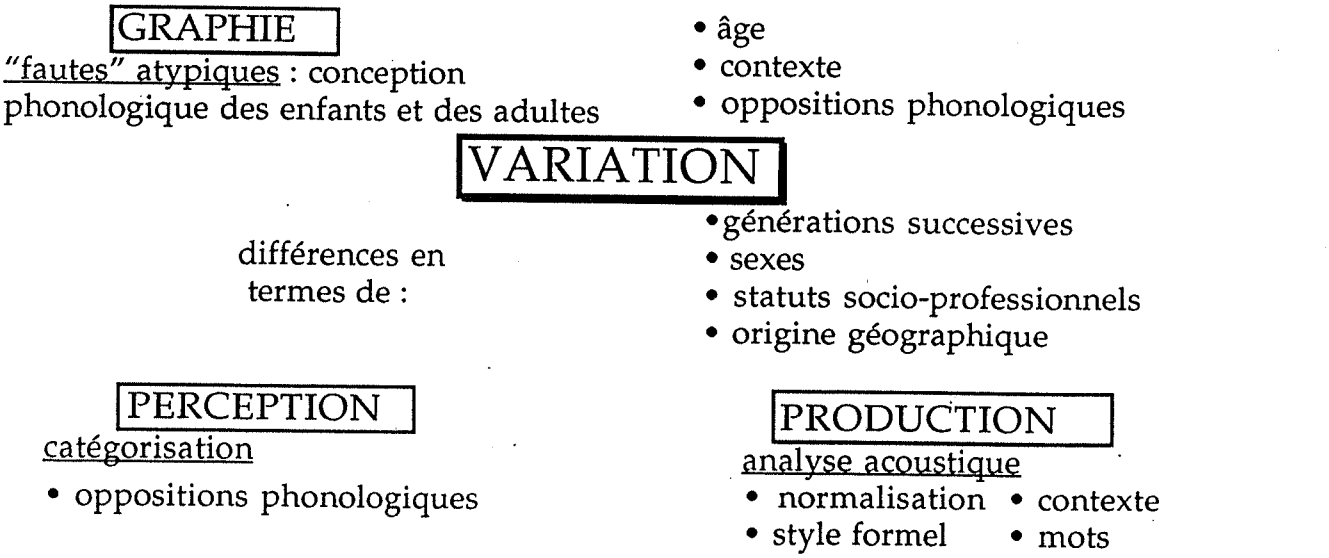
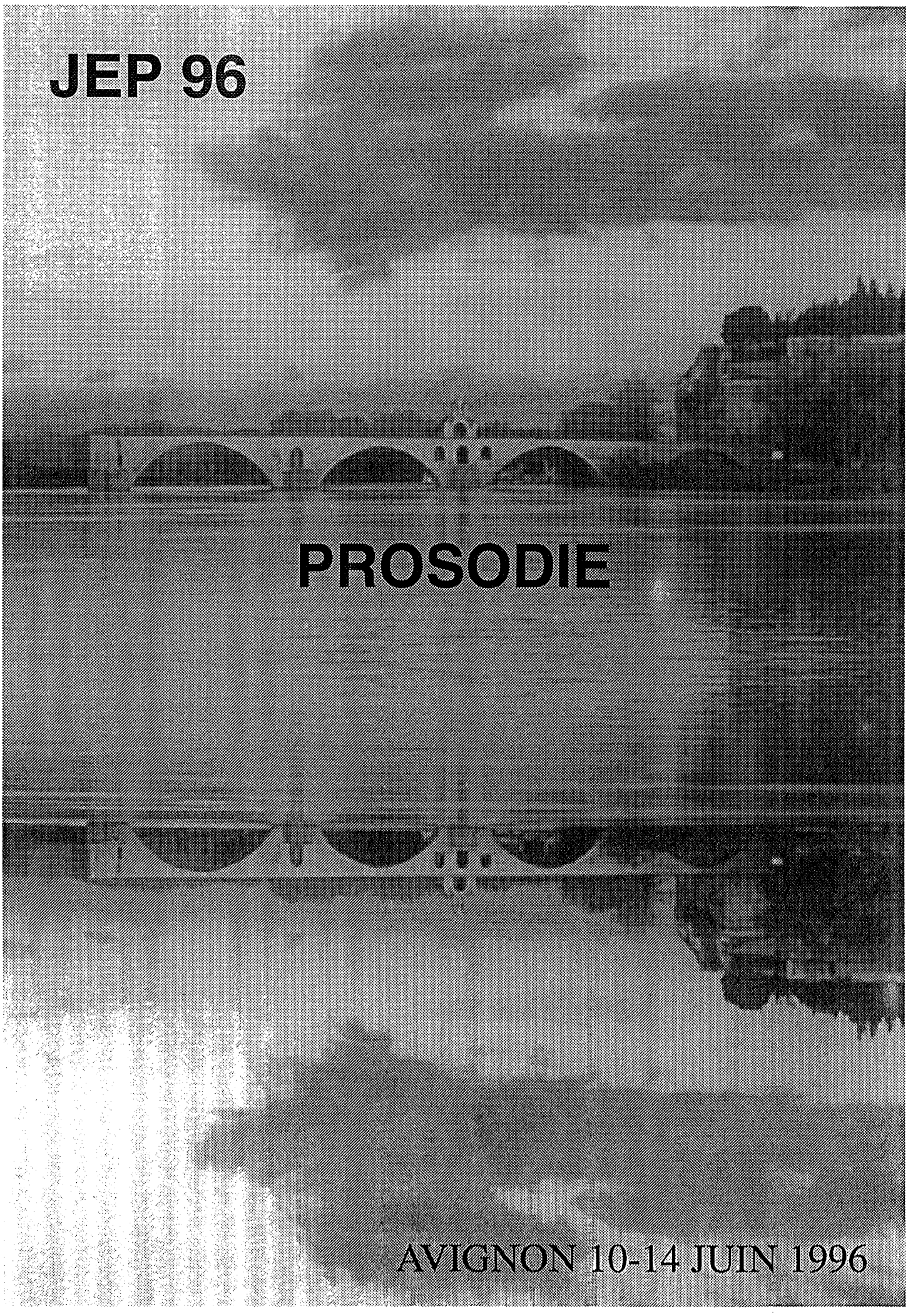


Figure 3 : Modèle tridimensionnel pour l'étude de la variation et des changements phonétiques en cours.

JEP 96



PROSODIE

AVIGNON 10-14 JUIN 1996

ACTES DE DIALOGUE ET PROTOTYPES MÉLODIQUES

Mariette Bessac, Nathalie Colineau, Geneviève Caelen-Haumont

Laboratoire de Communication Langagière et d'Interaction Personne-Système (CLIPS - IMAG)
Université Joseph Fourier - B.P. 53 ; F-38041 Grenoble Cedex 09 - FRANCE
tel : (+33) 76 51 45 10 ; fax : (+33) 76 44 66 75
e-mail : Mariette.Bessac@imag.fr ; Nathalie.Colineau@imag.fr ; Genevieve.Caelen@imag.fr

ABSTRACT

This paper concerns the study of natural dialogue in which the prosodic, linguistic and pragmatic components (reduced to melody, lexical units and dialogue acts) are particularly observed in order to show that a pragmatic analysis can be automatically performed from the melodic values and the lexical analysis.

Three analyses are performed (on the recording and transcription of about 90 minutes of conversation) to establish possible links: melodic prototypes are drawn from the pragmatic analysis and the prosodic measures. They give a standard pattern for every modality, which can be used to define the possible types and modalities of any utterance.

1. INTRODUCTION

Le travail présenté dans cet article porte sur le dialogue homme-homme naturel. Une série de dialogues oraux finalisés forment le corpus d'étude, qui est segmenté en actes, d'après la théorie des actes de langage. Pour cette étude, la théorie (trop monologique pour notre corpus) a dû être adaptée au dialogue, ce qui veut dire que les énoncés sont analysés en fonction de la situation d'énonciation et du but illocutoire du locuteur.

La compréhension des actes de dialogue requiert la contribution d'éléments à la fois linguistiques, situationnels et prosodiques. Cet article traite de l'importance de la prosodie dans la compréhension des actes de dialogue, et plus particulièrement de la mélodie : la mélodie permet-elle de discriminer les actes de dialogue entre eux ou permet-elle seulement d'aider à leur détermination ? Toutefois dans cette première étude nous restreindrons ce problème général au cas particulier des questions et à la forme générale de la courbe intonative.

Nous présenterons tout d'abord l'analyse pragmatique, préliminaire à l'étude prosodique, puis nous présenterons la méthode d'investigation et ses résultats.

2. RECUEIL DU CORPUS

Le corpus d'étude est un enregistrement audio de six conversations entre deux personnes (soit 12 locuteurs) qui simulent une conversation téléphonique entre un touriste et un employé de l'office du tourisme. Ils avaient deux tâches à réaliser en collaboration, à partir de connaissances différentes et de buts divergents (Gryl, 1995).

Le corpus a été enregistré en chambre sourde afin d'obtenir la meilleure qualité acoustique pour les mesures prosodiques. Les transcriptions orthographiques de ces corpus (Morel, 1989) servent de support au découpage en actes de dialogue.

3. ACTES DE LANGAGE ET ACTES DE DIALOGUE

Austin en 1962 (Austin, 1962) puis Searle en 1969 (Searle, 1969) considèrent tout énoncé réalisé comme un acte (ou une succession d'actes) qu'ils appellent acte de langage. Ils déterminent trois niveaux d'analyse dans un acte de langage :

- l'acte locutoire,
- l'acte illocutoire,
- l'acte perlocutoire.

L'acte illocutoire est celui qui nous intéresse ici. Il est défini à la fois par sa force, qui détermine ce qui est réalisé (requête, question, information, etc.) et par son contenu propositionnel (ce qui est demandé, donné en information, etc.).

L'étude présentée ici porte sur l'identification des éléments de la force illocutoire, c'est-à-dire des marqueurs du type d'acte.

3.1. Analyse en actes de dialogue

Il n'existe pas de relation biunivoque entre la forme des énoncés et les actes de dialogue correspondants. Par exemple, une question correspondant au type d'acte "question introduite" peut être énoncée de différentes manières :

- (1) "Quelle rue est-ce qu'il faut que je prenne ?"
- (2) "Quelle rue je prends ?"
- (3) "Je prends quelle rue ?"
- (4) "Quelle rue ?"
- (5) "Laquelle je prends ?"

L'énoncé produit n'aura d'ailleurs pas toujours une forme syntaxique de question.

Inversement, un énoncé peut correspondre à différents types d'actes, selon la situation d'énonciation : (6) "Il y a un pont suspendu" peut être une question ou une information donnée (avec ou sans nuance de surprise), selon l'intonation donnée par le locuteur.

Il existe donc nécessairement un ensemble de marques (situationnelles et prosodiques) qui s'ajoutent aux marques linguistiques et qui sont pertinentes à l'oral. Les marqueurs prosodiques ont été étudiés préférentiellement aux marqueurs situationnels car ils sont quantifiables et peuvent être facilement mesurés sur notre corpus.

La deuxième partie de cet article porte sur la présentation des marqueurs prosodiques.

4. ANALYSE PRAGMATIQUE

La prosodie est pertinente dans de nombreux actes et plus particulièrement dans les questions. En effet, elle est parfois la seule marque qui permette d'identifier l'acte de dialogue (cf. exemple 6).

4.1. Les types de questions du corpus

On a déterminé plusieurs groupes de questions dans le corpus, selon le type de réponses apportées.

4.1.1. Les questions appelant une complétion

- La question introduite : (7) "Vous appelez de quelle borne exactement ?"

4.1.2. Les questions appelant un accord (vs. un désaccord)

- La question oui/non : (8) "J'voudrais vous d'mander si vous avez un pont en haut à gauche ?"
- La demande de confirmation : (9) "Vous êtes à vélo, hein ?"
- La question alternative : (10) "Euh, du côté est ou du côté ouest ?"

4.1.3. Les questions dont l'énoncé littéral ne correspond pas à l'acte effectif

- La question oui/non indirecte : (11) "Et pour le restaurant traditionnel, vous avez des ordres de prix ? "
- La requête informative : (12) "Et j'aurais voulu aller au centre commercial."

Dans cet article, nous nous intéresserons aux questions introduites et aux questions oui/non.

4.2. L'étiquetage du corpus

Selon cette classification, toutes les questions du corpus ont été localisées et identifiées. Pour chaque type de question, on a donc défini trois étiquettes : début d'énoncé, début d'énoncé avec prise de parole (car la prise de parole risque de perturber les formes intonatives) et fin d'énoncé. La prise de parole (cf. (10), élément souligné) est un élément phatique qui sert uniquement à prendre ou à conserver le tour de parole. Elle peut être constituée d'un ou plusieurs mots (euh, eh ben, alors, etc.).

Toutes les questions du corpus ont été étiquetées à partir de notre classification, chaque étiquette indiquant le type d'acte et le type de frontière (début avec prise de parole, début ou fin d'énoncé).

5. PRINCIPE D'ANALYSE

L'interprétation des énoncés dépend de la situation d'énonciation et comme on l'a vu, de trois composantes : la linguistique, la prosodie et la pragmatique (Caelen-Haumont, 1996). Notre travail est d'examiner leurs relations deux à deux (Bessac et Caelen-Haumont, 1995 ; Colineau et Caelen, 1995). Cet article présente plus particulièrement la relation qui lie les actes de dialogue à la mélodie. Cette dernière résulte de deux constituants : la courbe générale liée à la forme syntaxique de l'énoncé et les perturbations locales liées au poids que le locuteur met sur chaque élément. Comme nous l'avons dit ci-dessus, nous limiterons notre analyse, dans cet article, à l'intonation des actes de dialogue.

On cherche à déterminer pour chaque acte au moins un prototype mélodique. La première hypothèse est donc relative à l'existence de ces prototypes : existe-t-il un ou plusieurs prototypes pour chaque type de question ?

Comme certains énoncés présentent peu (voire pas du tout) de marques syntaxiques, on peut imaginer que la prosodie pourrait compenser ce manque. Une seconde hypothèse, relative à un lien entre les prototypes trouvés et les regroupements syntaxiques possibles des énoncés associés, doit être vérifiée : existe-t-il une corrélation entre les éléments syntaxiques d'un énoncé et sa courbe mélodique ?

6. MÉTHODE D'ANALYSE DE LA PROSODIE

Pour chaque type de question, les courbes ont été réparties en trois groupes : les énoncés avec prise de parole (G1), les mêmes énoncés dont la prise de parole a été supprimée (G2) et enfin les énoncés sans prise de parole (G3). Les groupes G1 et G2 sont étudiés en parallèle pour savoir si la prise de parole fait partie de la question ou si elle la perturbe.

Les courbes de F0 ne représentant que les parties voisées du discours, elles sont discontinues et très accidentées. Elles ont besoin d'être comblées et lissées, ce qui entraîne inévitablement une distorsion. C'est pourquoi nous avons préféré travailler sur les

courbes mélodiques tracées à partir des noyaux vocaliques détectés sur le signal. Comme les énoncés du corpus sont de longueur variable, ils ont été ramenés à une même unité (100 points pris régulièrement sur chacune des courbes) pour pouvoir être comparés.

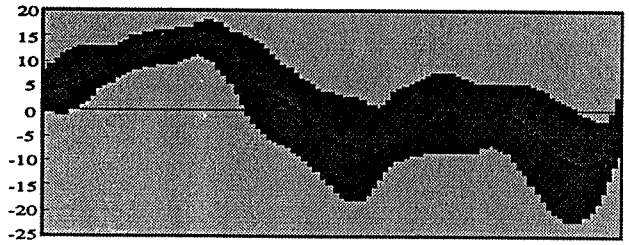
Pour chaque groupe, G1, G2 et G3, des questions introduites et des questions oui/non, une matrice de corrélations a été calculée. Les courbes les plus corrélées (Dagnelie, 1992) ont été regroupées pour former le ou les prototypes. La moyenne et l'écart type de chaque groupe servent à représenter le prototype du groupe (cf. Prototype A, paragraphe suivant).

Les prototypes obtenus ont de très bons scores comme on peut le voir sur le (tableau 1).

Tableau 1 : Résultats

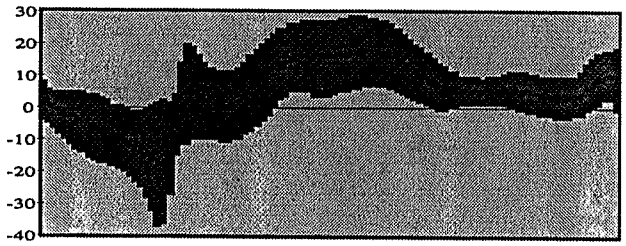
	Questions introduites	Questions oui/non
G 1	prototype A : 47,5% prototype B : 43%	prototype E : 46% prototype F : 31%
G 2	prototype C : 81%	prototype G : 38,5% prototype H : 27%
G 3	prototype D : 76%	prototype I : 46% prototype J : 44%

6.1. Les questions introduites

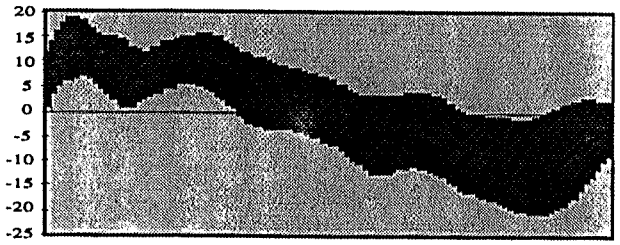


Prototype A : ■ moyenne, ■ écart type

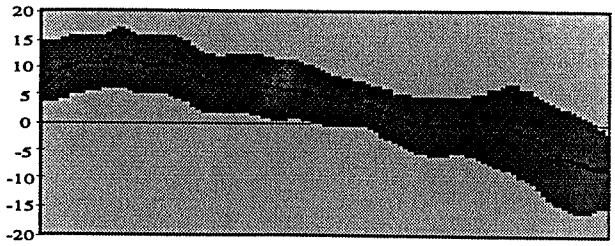
L'axe vertical est en 1/8ème de tons ; l'axe horizontal est temporel. Les prototypes sont plus ou moins représentatifs selon le nombre de courbes qu'ils représentent par rapport à l'ensemble des courbes de chaque modalité.



Prototype B



Prototype C



Prototype D

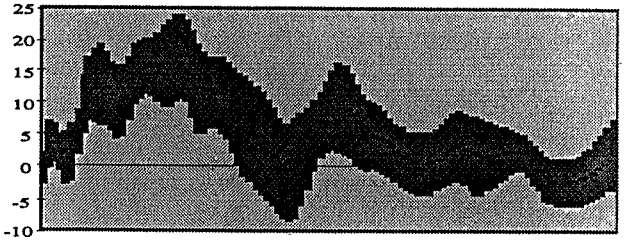
Il semble que les questions introduites sont caractérisées par une courbe descendante, excepté le prototype B. Ce phénomène peut s'expliquer par le fait qu'à l'exception des énoncés très courts, l'interrogation porte sur la fin ou sur la deuxième partie de l'énoncé.

On pourrait même penser que dans le cas où les questions sont marquées linguistiquement par la présence d'un interrogatif, la mélodie pourrait se rapprocher de celle d'une énonciation et être ainsi moins pertinente. Les marques linguistiques et prosodiques seraient alors réparties en distribution complémentaire.

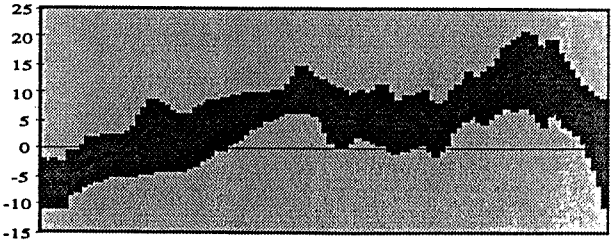
Pour cette recherche d'oppositions syntaxiques et/ou d'éléments linguistiques récurrents, une étude plus approfondie est en cours.

6.2. Les questions oui/non

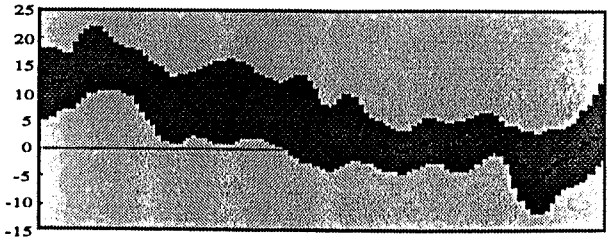
Les prototypes ont un écart-type moyen assez important, mais parfois réduit. Faute de place, le lieu exact de l'étranglement et l'unité lexicale associée ne sont pas détaillés dans cet article.



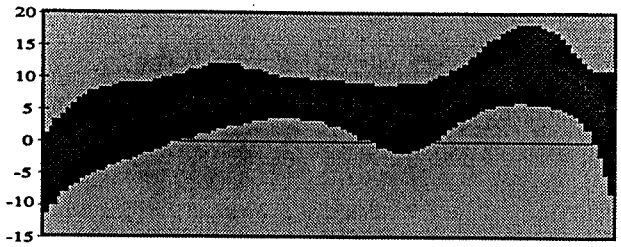
Prototype E



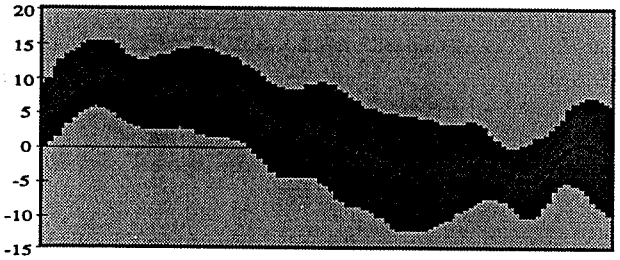
Prototype F



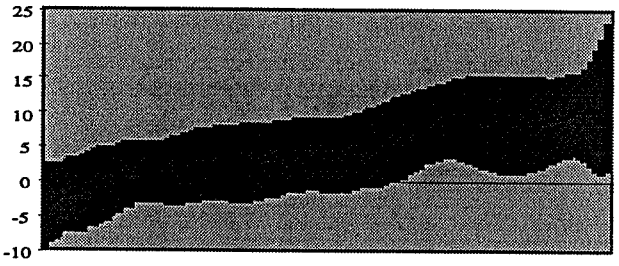
Prototype G



Prototype H



Prototype I



Prototype J

Pour les résultats des groupes G1 et G2 (qui représentent les mêmes énoncés avec et sans la prise de parole), on remarque des ressemblances de formes et de pentes entre les prototypes E et G d'une part et F et H d'autre part. De plus, ces prototypes regroupent presque les mêmes énoncés. On peut alors penser que la prise de parole ne perturbe pas la question et en fait partie, mais cela reste à confirmer sur les autres types de questions (demande de confirmation, question alternative, question oui/non indirecte et requête informative) et sur les autres types d'énoncés (déclarations, commandes, etc.).

Comme pour les questions introduites, des corrélations syntaxiques fines sont recherchées au sein de chaque prototype. On s'attend à trouver des regroupements de questions dans les énoncés des prototypes (commençant par "est-ce que", sans interrogatif, etc.). Cela nous permettrait, pour le cas des questions, de répondre positivement à la seconde hypothèse concernant une corrélation entre la présence de certains éléments linguistiques et la forme de la courbe mélodique associée.

En comparant groupe par groupe les prototypes des questions introduites avec ceux des questions oui/non, on constate que certains sont assez semblables (A et E, C et G, D et I), ce qui nous laisse penser que nous touchons là une limite possible de cette recherche, à savoir que les prototypes, si représentatifs qu'ils soient des types de questions, ne nous permettent pas de différencier une question introduite d'une question oui/non de façon certaine.

Nous supposons donc que la mélodie permet de déterminer les actes de dialogue produits (déclaration, question, commande, etc.) mais pas de les différencier entre eux (c'est-à-dire d'identifier le type de déclaration, de question, de commande, etc.). Mais cela reste encore à vérifier sur le reste des énoncés.

7. CONCLUSION

Nous pouvons à présent répondre à nos hypothèses de départ : il existe bien un ou plusieurs prototype(s) mélodique(s) pour chaque type de question, mais les résultats actuels ne permettent pas de savoir si un lien plus direct ou plus fin existe entre les événements mélodiques et la syntaxe des énoncés.

La prise de parole semble faire partie de la question, tout au moins pour les deux types de questions étudiés. Cette observation demande à être confirmée par la suite sur les autres types de questions et sur les autres types d'énoncés.

Il reste à confirmer la ressemblance des prototypes présentés ici avec ceux des autres types de questions. Si les prototypes de toutes les questions se ressemblent, on ne pourra pas déterminer le type d'acte d'après la forme mélodique globale.

Ils devront également être comparés avec ceux des autres actes de dialogue pour savoir quels types d'énoncés la prosodie permet de reconnaître et quel degré de confiance on peut leur accorder.

8. BIBLIOGRAPHIE

- Austin J. L. (1962) *Quand dire c'est* faire, Seuil, 1970
- Bessac M. et Caelen-Haumont G. (1995) Analyses pragmatique, prosodique et lexicale d'un corpus de dialogue oral homme-homme, *III*^{èmes} *journées internationales d'analyse de données textuelles*, 363-370
- Caelen-Haumont G. (1996) *Prosodie et sens*, ouvrage soumis à publication
- Colineau N. et Caelen J. (1995) Étude de marqueurs dialogiques dans un corpus de conception, *Le communicationnel pour concevoir*, 203-222
- Dagnelie P. (1992) *Statistique théorique et appliquée*, Presses agronomiques de Gembloux
- Gryl A. (1994) Analyse cognitive des descriptions d'itinéraires, *Actes du premier colloque jeunes chercheurs en sciences cognitives*, 211-220
- Morel M.-A. (Resp.) (1989) *Analyse linguistique d'un corpus*. Publications de la Sorbonne Nouvelle
- Searle J. R. (1969) *Du cerveau au* savoir, Hermann, 1985

APPROCHE PROSODIQUE ET PRAGMATIQUE DES MODULATIONS

Roxane BERTRAND*, Florence CASOLARI**

* Institut de phonétique d'Aix-en-Provence. Laboratoire Parole et Langage URA 261, CNRS.

Tel.: 42 95 36 24 - Fax: 42 59 50 96 - e-mail: roxane.bertrand@lpl.univ-aix.fr

** Département de Linguistique générale d'Aix-en-Provence - 29 avenue R. Schuman. 13621 Aix-en-Provence.

Tel: 42 95 35 95 - Fax: 42 20 48 80

ABSTRACT

Prosodic features are examined to establish a formal and functional typology. They are studied within the larger pragmatic frame of "modulations": a phenomenon specific to verbal interactions. Evidence of the prosodic organization of these modulations is given here. It assesses the presence of this notion in pragmatics and furthermore reveals some of its functional aspects. Finally, results give support to the presence of transition points called "pivot" within the considered structure.

1. INTRODUCTION

Notre travail s'inscrit dans une double perspective, pragmatique et prosodique, qui cherche à caractériser, dans ses aspects formels et fonctionnels, le phénomène appelé modulation (Vion, 1992). D'un point de vue pragmatique, les modulations sont des mouvements[1] interactifs caractérisés par une modification d'attitude du locuteur par rapport au contenu de son propre discours. L'intérêt de l'étude consiste à mettre en relation ces considérations d'ordre pragmatique avec l'organisation prosodique de la séquence.

Les modulations étudiées sont extraites d'un corpus de 23 interactions (corpus *AFL*[2]), enregistrées audiovisuellement, dans lesquelles deux étudiants discutent 10 minutes sur un thème choisi entre 3 sujets.

2. OBJET DE L'ETUDE

Notre analyse porte sur des structures simples et assez brèves (dans le même tour de parole) bien qu'une courte intervention (ou réaction) de l'interlocuteur aient pu les susciter. Nous sommes conscient que d'autres procédés de modulations existent, qu'il serait intéressant d'observer par ailleurs.

Les modulations que nous nous attachons à décrire sont caractérisées par deux phases distinctes: la phase de tension (T) et la phase de modulation proprement dite (M).

EX [c'est pas qu'c'est rétrogade] mais enfin #
T
[c'est un peu dépassé]
M

T est souvent marquée par un terme plus ou moins explicite, une expression spécifique, dont le locuteur pense qu'ils peuvent entrainer un certain malaise chez son interlocuteur. Une utilisation lexicale jugée inadéquate, exagérée... par le locuteur en T ("rétrogade"), entraine un réajustement immédiat, sous la forme d'une correction, d'une reformulation, d'une atténuation..., qui correspond à la phase M (tend à préserver le bon déroulement de l'interaction). M est introduite par un connecteur (31/37 cas), dont 28 "enfin" (isolé ou combiné). Dans l'exemple cité, nous isolons les connecteurs par rapport à T et M, du fait de leur rôle de "disjonction" pragmatique, qui, selon notre hypothèse, est également liée à une disjonction prosodique.

Un certain nombre de modulations sont énoncées sous forme d'incise. Nous avons donc mis en relation quelques-uns de nos résultats prosodiques avec ceux obtenus sur ce phénomène.

Enfin, nous souhaitons rattacher les modulations à deux concepts centraux dans l'analyse pragmatique, et notamment dans celle des interactions.

Le premier est lié au *principe de coopération*, (fondé sur 4 maximes) (Grice, 1979), essentiel dans la gestion de toute interaction. Ainsi, dans notre corpus, les participants doivent coopérer d'autant plus que leur contrat de parole (contraintes codifiant les pratiques socio-langagières d'une situation donnée) (Charaudeau, 1983) est limité et de caractère fermé: ils doivent en effet "tenir" une discussion plutôt impliquée (présentation d'opinions personnelles) avec un inconnu, risquant chacun à tout moment de faire "mauvaise figure" (Goffman, 1974).

[1] "move" (Goffman, 1973), unité d'analyse de type verbal, non verbal, para-verbal, qui a une fonction distinctive dans un ensemble de circontances de communication.

[2] Analyse des Fonctionnements Langagiers.

D'autre part, le *phénomène polyphonique* (Ducrot, 1980,) se manifeste dans les modulations où un locuteur peut mettre en scène, dans un même énoncé, plusieurs énonciateurs.

3. PRINCIPES D'ANALYSE PROSODIQUE

Les phases T, M + connecteur caractérisent nos modulations. Toutefois, il nous semble utile de définir des critères d'identification plus précis. La mise en évidence d'un schéma prosodique spécifique des modulations semble justifiée pour aboutir à une typologie. En outre, on peut faire l'hypothèse que c'est grâce à ce schéma prosodique particulier que l'interlocuteur repère rapidement la modulation.

Nous avons observé d'un point de vue prosodique 20 cas de modulations, dans le but de dégager des caractéristiques prosodiques récurrentes qui en permettent l'identification.

Le paramètre de hauteur, à travers l'évolution et les variations de la courbe de F0, constitue le premier paramètre analysé. Selon notre hypothèse, un contour intonatif spécifique récurrent, malgré les différences individuelles et le matériau verbal différent, doit affecter les modulations, et permettre de distinguer T de M.

Dans un second temps, nous considérons les paramètres temporels suivants: les pauses silencieuses, qui participent des traits intonatifs en tant que planificateurs des constituants du discours (Rossi, 1980), les allongements syllabiques, qui constituent un des critères d'accentuation de la langue, -dans la typologie accentuelle de Di Cristo (1981), l'accent *primaire* se caractérise par un ton haut et un allongement-, et les "euh" d'hésitation. Nous les observons ici dans le contexte d'apparition précis qui est le lieu de disjonction, actualisée par le connecteur entre T et M. Dans ce cadre, nous les considérons en tant qu'indices catégoriels uniquement, en terme d'absence ou de présence, une étude quantitative de ces phénomènes étant en cours par ailleurs. Nous espérons déterminer leur éventuelle fonction (différenciatrice?) en ce lieu de disjonction.

Enfin, nous mesurons la durée de T et de M (en ms), nous calculons leur vitesse d'élocution (en syllabes par secondes), ainsi que le nombre de syllabes qui les composent.

4. RESULTATS

4.1 Contour intonatif

Nous constatons, dans 90% des cas, un contour intonatif à l'allure générale suivante:
- une montée intonative finale abrupte sur la dernière ou l'avant dernière syllabe de T.
- une montée intonative souvent plus progressive sur la partie finale de M.
- une chute intonative qui affecte le connecteur présent (voir figure 1).

Les deux premiers résultats confirment les résultats obtenus pour les incises (Delomier et Morel, 1986; Achez, 1995), certaines modulations étant produites en incises. En revanche, M, (incise ou pas), n'est jamais énoncée à un niveau tonal plus bas que la proposition insérante, ce qui infirme les résultats de certains de ces mêmes travaux.

4.2 Paramètres présents entre T et M:

- un connecteur: 16 cas sur 20 au total
-une pause silencieuse: 15/20, (dont 10: connecteur + pause).
-allongement syllabique final en T dans 11 cas, dont 6 "euh" d'hésitation; (dans 10 cas, allongement+pause).

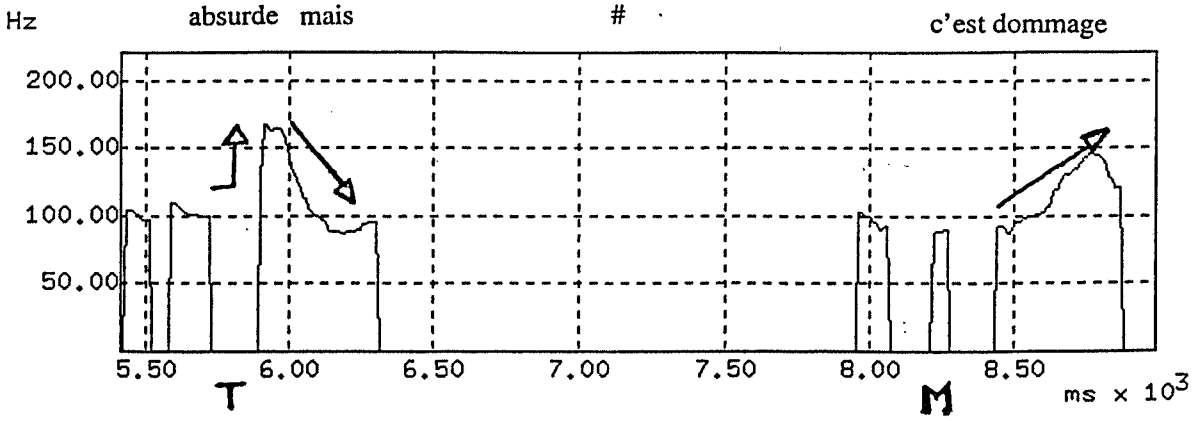


Figure 1: courbe illustrant un exemple de modulation:. Montée intonative finale en T, chute mélodique sur le connecteur (mais), pause silencieuse. M: montée intonative finale

4.3 Vitesse d'élocution:

Dans tous les cas, elle varie entre T et M (comme avant et pendant l'incise). On peut donc parler d'opposition entre les deux phases, mais les données actuelles (qui nécessitent d'être enrichies) ne permettent pas de donner de tendance précise (accélération en M et ralentissement en T par exemple).

Enfin, T et M, comportent à peu près le même nombre de syllabes, ou bien sont d'une durée (en ms) relativement proche.

5. COMMENTAIRE

Cette étude est conçue comme le point de départ d'un travail qui tente de lier les aspects pragmatiques et prosodiques, et qui a pour objectif premier la mise en évidence d'un procédé linguistique spécifique, jamais envisagé expérimentalement. L'exposé de ces résultats préliminaires montre certaines tendances, qui permettent d'aboutir à des premières conclusions encourageantes, et qui, de ce fait, justifie des études ultérieures, fondées sur davantage de données et prenant en compte les aspects quantitatifs des critères prosodiques retenus. Ici, en effet, les indices prosodiques étudiés (sens de la courbe de F0 par exemple) ont été envisagés en tant qu'indices "signifiants" d'un point de vue pragmatique.

Les données intonatives et le relevé de certains événements temporels nous incitent à accorder une place essentielle à la transition entre T et M, qui constitue, semble t-il, la plaque tournante de la séquence. Nos conclusions concernent donc essentiellement la façon dont les deux phases s'articulent l'une par rapport à l'autre, et l'interprétation, en terme pragmatique, des évènements prosodiques de cette articulation.

Le pivot: unité formelle?

La présence d'un connecteur associé à des pauses, des allongements finaux ou des euh d'hésitations marque la modulation et illustre la transition entre T et M, transition soulignée par des évènements récurrents de type intonatif.

La mise en évidence de ces divers éléments nous incite à parler de pivot, qui renvoie dans son sens littéral, à un point autour duquel on observe des changements (de direction notamment). On peut établir des frontières entre T, le pivot et M, mais ces frontières sont-elles fondées sur des critères formels permettant d'établir l'existence d'une véritable unité prosodique? Ainsi, si le critère de délimitation choisi pour déterminer la transition (à gauche) est le début de la chute mélodique, on observe que celle-ci commence soit sur la dernière syllabe de T (pendant l'allongement par exemple), soit sur le "euh" qui suit parfois cette syllabe, ou bien directement sur le connecteur. A droite, la reprise du niveau tonal de T (avant le pic haut final) en M, constitue notre critère de délimitation. Or, autour de ces événements mélodiques, nous notons dans certains cas la présence de pauses silencieuses. De ce fait, à ce stade de l'étude, nous ne pouvons affirmer que le pivot est fondé sur un critère (intonatif) unique de délimitation, ses frontières étant constituées d'un ensemble d'évènements nécessaires (chute intonative, pause, niveau tonal spécifique, connecteur) et facultatifs (allongements syllabiques, euh).

Le pivot: aspect fonctionnel

La rupture constatée entre T et M dépend, entre autres, de critères intonatifs (changement de direction de pente) et temporels (pauses, allongements, variation de débit). En parole spontanée, la pause silencieuse constitue une borne entre les groupements; elle peut alors, seule, constituer une transition entre T et M. On fait ainsi l'hypothèse que la pause, associée à la chute mélodique, constitue un fort indice de démarcation, d'autant plus fort si le connecteur y est associé. Le pivot assume donc une première *fonction démarcative*.

A travers les connecteurs, qui sont les marques formelles d'une opération de nature métadiscursive (Vion et Giacomi, 1987), et le schéma intonatif "montant-descendant" observé (fin de T, début du pivot), le pivot assume aussi une *fonction discursive*.

En outre, ce type de configuration mélodique assume un rôle non seulement discursif mais aussi informatif. La montée intonative (une caractéristique accentuelle connue) a pour fonction, selon nous, la mise en relief de l'élément final de T, qui porte (sémantiquement) le poids de tension le plus fort de la séquence ("idiot", "absurde"...). Après cet élément accentué, la présence du connecteur et des évènements prosodiques relevés, signalerait la volonté de modification de l'attitude du locuteur, la chute intonative jouant un rôle essentiel.

Les modifications de pente (vers le haut comme vers le bas) dans la courbe de F0, traduisent un accent (voir notamment Guaïtella, 1992). Par cette chute mélodique, le locuteur souhaiterait attirer l'attention de l'interlocuteur. Ce lieu de transition, à forte valeur informative, signifierait :"attention, je vais dire quelque chose d'important", (la chute manifestant cette position plus modérée par rapport à la montée, plus souvent signe d'appel, de relance).

Nous retenons également la *fonction discursive stylistique*, que Freland (1995)

attribue à la pause et que nous appliquons, ici, au pivot. Cette fonction permet en effet de brusques changements de sujets, qui marquent les unités du discours.

La rupture intonative et temporelle est donc accompagnée d'une transition dans l'argumentation (illustrée par le connecteur) mais aussi et enfin dans l'interaction. En effet, ce lieu représente le noyau dur de l'interaction dans le sens où il marque le moment privilégié où le locuteur prend conscience de la présence de l'autre, puisqu'il modifie son opinion, du moins sa façon de l'exprimer. C'est là qu'il rétablit l'équilibre entre ce qu'il vient de dire et ce qu'il faut dire pour préserver non seulement sa face, mais aussi celle de l'autre (en atténuant sa position précédante plus virulente), et garantir la suite de l'interaction. Le pivot assume donc, aussi, une *fonction interactive*.

6. CONCLUSION

Les résultats de cette étude préliminaire permettent de mettre en évidence un schéma prosodique spécifique des modulations. Ceci assigne une existence réelle à ce procédé pragmatique, jamais vérifié expérimentalement, mais dont les locuteurs ont conscience, du moins qu'ils attestent, puisqu'ils utilisent les mêmes procédés prosodiques dans des séquences linguistiques diverses.

Nos résultats permettent aussi d'illustrer les deux grandes fonctions que l'on reconnait traditionnellement à la prosodie. La fonction de *structuration*, qui permet de faire des groupements et la fonction de *mise en relief* d'un élément.

Enfin, nous mettons en évidence un lieu de transition, le pivot, établi sur la base d'indices prosodiques pour l'essentiel, qui assume principalement les diverses fonctions prosodiques, discursives, informatives et interactives de la séquence étudiée. Par la suite, nous nous attacherons à la description très précise de ce pivot, en terme quantitatif notamment (étude des durées) afin de confirmer certaines de nos hypothèses à propos de son éventuel statut d'entité (voire d'"unité") prosodique.

En outre, une analyse est en cours sur l'aspect plus linguistique de ces modulations ("stratégies linguistiques") tandis qu'une description de type gestuel est également prévue.

7. BIBLIOGRAPHIE

Achez C. (1995) *L'identification des incises*, Mémoire de Maîtrise, Université de Provence.

Charaudeau P. (1983), *Langage et discours; Eléments de Sémiolinguistique*, Hachette, Paris.

Delomier D. et A.M. Morel (1986) Caractéristiques intonatives et syntaxiques des incises, *DRALV* 34-35, 141-160.

Di Cristo, A. et D. Hirst (1993) "Rythme syllabique, rythme mélodique et représentation hiérarchique de la prosodie du français", *Travaux de l'Institut d'Aix-en-Provence*, 15, 9-24.

Ducrot O. (1980) *Les mots du discours*, Edition de Minuit.

Freland-Ricard M. (1995) *Analyse multilingue des erreurs prosodiques chez l'apprenant étranger. Essai de paramétrisation*, Thèse de Doctorat en Phonétique, Université de Provence.

Goffman E. (1974) *Les rites d'interaction*, Ed. de Minuit.

Grice P. (1979) in "La conversation", Logique et Conversation, *Communication* n° 30.

Guaïtella, I. (1992) Hésitations vocales en parole spontanée: réalisations acoustiques et fonctions rythmiques, *Travaux de l'Institut d'Aix-en-Provence*, 14, 109-130.

Rossi M (1980) Le Français, Langue sans accent, L'accent en français contemporain, *Studia Phonetica*, 15, Didier, Paris, Montréal, 13-52.

Vion R. (1992) *La communication verbale: Analyse des Interactions*, Hachette, Paris.

Vion R. et Giacomi A., (à paraître): "Connecteurs et mise en place des activités discursives" in Acquisition du français par des travailleurs marocains, *Papiers de Travail n°3 du GRAL-AIX*, Publications de l'Université de Provence

DE LA RELATION ENTRE LE TIMING DES MOUVEMENTS MÉLODIQUES ET L'ACCENTUATION DES SYLLABES

Frédéric Beaugendre, Dik J. Hermes

Institute for Perception Research (IPO)
5600 MB Eindhoven, The Netherlands
E-mail : beaugend@natlab.research.philips.com

ABSTRACT

In this study is investigated the relation between the timing of a rising or a falling pitch movement and the syllable it accents. Recent experiments run with Dutch subjects have shown that the percept of accentuation is induced by a change in pitch at the start of the movement.

Here, two different experiments are described. First of all, in order to be able to locate the accentuation boundary for syllables other than the /ma/ syllables used in previous studies, the phonemic constituents of the syllable are varied. The results confirm that the percept of accentuation is induced by a change in pitch at the start of the movement. Moreover, if this change of pitch is situated in a silent region, an interval of uncertainty arises in which the subjects either indicate the syllable preceding or the syllable following the silence as accented.

The second experiment, run with French native subjects, was carried out to check whether language specific information is used in the perception of accent. For the rising movements, it clearly appeared that the subjects are well able to do the task. The results show that the accentuation boundary is located about 50ms earlier than for the Dutch subjects. For the stimuli with the falling movements, most of the French subjects encounter serious problems to locate the accented syllable, and no consistency can be found in the results among subjects.

1. INTRODUCTION

Pour des langues telles que le hollandais et l'anglais, plusieurs types de mouvements mélodiques montants et descendants permettant d'accentuer une syllabe particulière sont inventoriés : plusieurs études distinguent deux types différents de mouvements montants ; le premier baptisé "early rise" commence avant l'attaque vocalique, le second "late rise" commence après l'attaque vocalique ('t Hart et al., 1991; Hill & Reid, 1977) ; des résultats similaires sont obtenus pour le français (Beaugendre, 1994). En revanche, cette distinction est moins claire pour les mouvements descendants. Ainsi, une seule catégorie phonétique est inventoriée pour le hollandais ('t Hart et al., 1991).

De récentes expériences ont été menées afin d'établir la relation existant entre le timing des mouvements mélodiques et la perception d'accentuation des syllabes (Hermes, 1995). Les résultats de ces expériences montrent que la notion d'accentuation est induite par une variation mélodique au tout début du mouvement.

Dans cet article, deux différentes expériences sont décrites. Tout d'abord, afin d'être capable de localiser précisément la région de situation du mouvement mélodique où se fait la transition de la perception de l'accentuation d'une syllabe à la syllabe suivante (nous appelerons désormais cette région la "frontière d'accentuation"), plusieurs catégories de syllabes doivent être envisagées.

La seconde expérience consiste à vérifier si cette relation entre le timing du mouvement et

la perception de l'accent est dépendante de la langue.

2. EXPÉRIENCE 1

2.1. Stimuli

Pour l'expérience déjà décrite par D.Hermes, la syllabe /ma/ fut utilisée. Dans cette expérience, la suite de cinq syllabes /mamamamama/ forme la phrase utilisée pour construire les stimuli. L'enregistrement original est la phrase /mamama/ à partir de laquelle la syllabe du milieu est dupliquée deux fois à l'aide de la technique d'analyse-synthèse TD-PSOLA (Hamon et al., 1989).

Toujours à partir de cette technique, les stimuli de cinq syllabes sont générés en superposant un mouvement mélodique montant ou descendant. Trois durées de mouvements sont utilisées : 80ms, 120ms et 160ms.

Tous les stimuli sont construits en décalant le mouvement considéré d'un intervalle de 20ms. Ces mouvements mélodiques sont superposés à une ligne de déclinaison fixée à 0.7 E/s (E est l'unité de l'échelle "Erb-rate"). La taille du mouvement mélodique est fixée à 1.5 E, soit environ 8 DemiTons.

Dans cette nouvelle expérience, la consonne n'est plus un /m/ mais un silence. Les stimuli seront donc formés à partir de la phrase /a.a.a.a.a/ dans laquelle /./ indique un silence.

La mise au point des stimuli respecte la même procédure. Cependant, afin de limiter le nombre de stimuli, tous les stimuli sont calculés en faisant varier la place du mouvement à partir du moment où la fin de celui-ci dépasse le milieu de la seconde voyelle jusqu'à ce que le début de ce mouvement se trouve avant le milieu de la quatrième voyelle. Les transitions 2-3, 3-4 seront donc observées.

L'ordre de présentation de tous les stimuli (256) est aléatoire. L'ensemble des stimuli fut présenté deux fois à chaque sujet. Chaque test commençait par dix stimuli non comptabilisés dans les résultats.

Afin de mesurer l'influence de la longueur du silence sur la localisation de la frontière d'accentuation, la même expérience fut menée en doublant, puis en diminuant de moitié la durée des silences dans le stimuli original.

2.2. Sujets et procédure

Quinze sujets hollandais furent utilisés pour cette expérience. Ils étaient installés dans une pièce calme. Ils écoutaient au casque les stimuli directement numérisés. L'environnement de test permet d'écouter les stimuli autant de fois que souhaité. La réponse est un choix forcé où le sujet doit indiquer quelle syllabe il a perçue comme la plus accentuée.

2.3. Résultats

Pour les trois expériences réalisées, les distributions furent calculées en fonction de l'instant de départ du mouvement mélodique considéré. Le moment que nous cherchons à déterminer est le moment où la sensation d'accentuation passe d'une syllabe à la suivante. A ce moment, les distributions vont se croiser. Le point de croisement des distributions est calculé après lissage à l'aide d'une fenêtre de Hanning à cinq points. Ce point est la frontière d'accentuation.

Les résultats pour les trois expériences apparaissent dans le tableau 2. Plusieurs remarques peuvent être faites : tout d'abord, comme pour les expériences précédentes, il apparaît que la notion d'accentuation est liée à la perception d'une variation mélodique au tout début du mouvement. En effet, lorsque l'on choisit les distributions en fonction de l'instant de départ de la variation mélodique, le point de rencontre des distributions varie peu avec la taille du mouvement considéré. D'autre part, cette faible variation semble indiquer que c'est en fait la différence de variation mélodique à partir du début du mouvement qui constitue le corrélat acoustique de la sensation d'accentuation (de l'ordre de 2 DemiTons). La seconde remarque concerne la différence observée entre les mouvements montants et descendants qui n'existait pas pour l'expérience précédente (Hermes, 1995) : il existe en fait un intervalle d'incertitude durant lequel les sujets sont plus difficilement capables de déterminer quelle syllabe est accentuée. Cette période est environ de la durée du silence qui sépare les deux voyelles consécutives. Ceci indique clairement que la mélodie est perçue durant une consonne voisée et que ce trait de

voisement est déterminant sur la localisation de l'accent. Cependant, il est interressant de remarquer que les sujets, durant cette période d'incertitude, répondent plus volontier la syllabe suivante lorsque le mouvement est montant et la syllabe précédente lorsque le mouvement est descendant.

3. EXPÉRIENCE 2

La seconde expérience menée visait à tester si ces résultats de perception sont dépendants de la langue considérée. Les mêmes expériences menées pour des sujets hollandais (avec les stimuli /mamamamama/ et /a.a.a.a.a/) furent donc menées avec des sujets français cette fois.

Afin de limiter le nombre de stimuli, nous nous sommes focalisés sur la frontière d'accentuation située entre les syllabes 3 et 4.

3.1. Stimuli, sujets et procédure

Les stimuli, le matériel utilisé et les conditions de tests étaient strictement les mêmes que pour la première expérience.

3.2. Résultats

Les figures 1 et 2 montrent un exemple des distributions obtenues pour la syllabe /ma/ (/m/ de durée normale, mouvements montants et descendants). Le tableau 2 présente la comparaison des résultats obtenus pour les sujets français et hollandais. En ce qui concerne les mouvements mélodiques montants, les résultats obtenus montrent que les sujets français sont capables de réaliser la tache demandée. Cependant, pour chacun des stimuli, la frontière d'accentuation est systématiquement située environ 60ms plus tôt que pour les sujets hollandais pour la syllabe /ma/ et 30ms plus tôt pour la syllabe /.a/. Ces résultats suggèrent deux explications possibles : les sujets français semblent avoir besoin d'indices plus clairs pour être capables de percevoir une syllabe comme accentuée; en outre, il faut noter que le mouvement montant arrivant tard dans la syllabe "late rise", très courant en hollandais, n'est rencontré que rarement en français (où il a uniquement une fonction expressive).

Pour les mouvements descendants, la plupart des sujets français ne sont pas capables de réaliser le test ; les résultats sont donc impossible à interpréter. Ce dernier point constitue une différence importante entre les deux langues qui peut être argumentée par le fait qu'en français, les mouvements mélodiques descendants ne sont pas définis comme des mouvements jouant un rôle déterminant dans la structure accentuelle (Vaissière, 1980; Beaugendre, 1994).

4. BIBLIOGRAPHIE

Beaugendre F. (1994) *Une étude perceptive de l'intonation du français*, thèse de doctorat, Université de Paris-Sud, Orsay.

Hamon C., Moulines E. & Charpentier F. (1989) A diphone synthesis system based on time domain prosodic modifications of speech, *ICASSP 89*, 238-241.

Hermes D.J. (1995) *Timing of pitch movements and acccentuation of syllables*, Annual progress report, Institut for Perception research.

Hill D.R. & Reid N.A. (1977) An experiment on the perception of intonation features; *International Studies on Machines Studies*, n°9, 337-347.

't Hart J., Collier R. & Cohen A. (1991) *A percepual study of intonation: an experimental-phonetic approach to speech melody*, Cambridge University Press.

Vaissière J. (1980) La structuration acoustique de la phrase française, *Annali 530-560, Scuola Normale Superiore Di Pisa.*

REMERCIEMENTS

Je tiens à remercier mes sujets hollandais de l'Institute for Perception Research et mes sujets français du LIMSI-CNRS pour leur patience et le temps qu'ils m'ont accordés. Ce travail est financé par le projet HCM de la Communauté Européenne.

Syllable	2		3		4	
Stimuli	C	V	C	V	C	V
/ma/	210	290	470	550	730	810
/.a/	202	285	555	640	910	995

Table 1 : Valeurs des débuts de segments pour les stimuli considérés, en ms (durée normale pour /m/ et /./) ; C=Consonne, V = Voyelle.

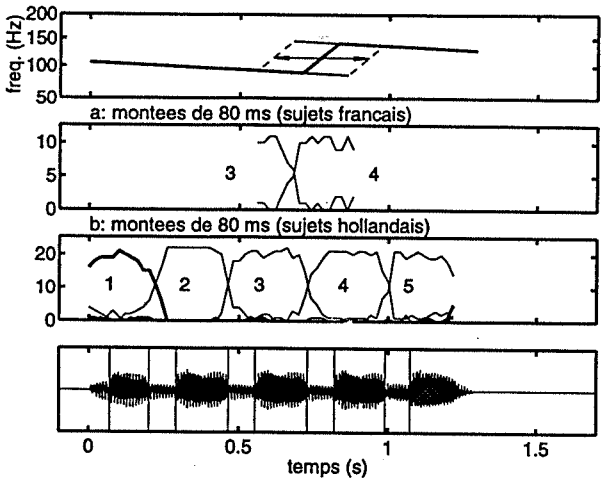


Figure 1: Distribution du début des mouvements mélodiques montants perçus comme accentuant les syllabes correspondantes (stimuli formés avec la syllabe /ma/, durée normale du /m/). Sont présentés de haut en bas : la situation du mouvement mélodique dans le stimuli pour les sujets français, les résultats pour les sujets français puis hollandais, le signal et les frontières phonémiques.

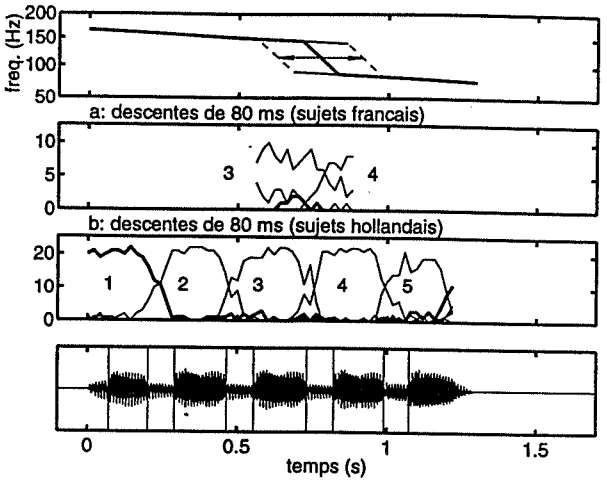


Figure 2: Distribution du début des mouvements mélodiques descendants perçus comme accentuant les syllabes correspondantes (stimuli formés avec la syllabe /ma/, durée normale du /m/). Sont présentés de haut en bas : la situation du mouvement mélodique dans le stimuli pour les sujets français, les résultats pour les sujets français puis hollandais, le signal et les frontières phonémiques.

Catégorie de la Syllabe	Langue	Durée du mouvement	Montées		Descentes	
			2-3	3-4	2-3	3-4
/ma/	français	160	-	674	-	-
		120	-	661	-	-
		80	-	645	-	-
	hollandais	160	459	740	452	706
		120	452	724	475	714
		80	464	734	477	748
/a/	français	160	-	821	-	-
		120	-	813	-	-
		80	-	795	-	-
	hollandais	160	515	854	605	907
		120	510	850	583	879
		80	495	836	561	856

Table 2 : Position de la frontière d'accentuation dans le stimuli en fonction de l'instant de début du mouvement (en ms), ceci pour des durées normales du /m/ et du silence /./ (voir tableau 2 pour une comparaison avec les frontières de segments). Pour les sujets français, la frontière entre les syllabes 2 et 3 n'a pas été étudiée, et les résultats pour les mouvements descendants ne sont pas significatifs.

ETUDE CONTRASTIVE DES PATRONS INTONATIFS EN ESPAGNOL ET EN ARAGONAIS - TRAITEMENT DE LA Fo

C. Franchon Cabrera, A. Rhardisse

Centre de Dialectologie de Grenoble - ICP/INPG/ENSERG
Université Stendhal - Grenoble III - BP 25 - 38040 Grenoble cedex 9
Tél.: 76 82 43 80 - Fax: 76 82 43 56

ABSTRACT

If we compare prosodic studies for ibero-romanic languages with other aspects of phonetics, we observe that there have been fewer such studies than in other fields of phonetics. We present here an intonative study for Spanish and Aragonese. Our aim is to study the principal features of the prosodic structuration for these two languages in relation to variations of the modality (assertive sentences vs interrogative sentences -total question-) while taking into account the main key points where we observe the decisive differenciation marks between these two langages.

1. INTRODUCTION

Nous situant dans l'une des traditions de la phonétique relative à l'étude de la substance acoustique, notre approche, de type instrumentale, prend en compte le processus de la production. Cette orientation vise à caractériser la phrase et ses composants majeurs (phrase, syntagmes, unités accentuelles, syllabes) par rapport aux paramètres physiques classiquement étudiés que sont, la variation temporelle de la fréquence laryngienne, la durée et l'intensité .

2. METHODOLOGIE

Une démarche commune dans la constitution du corpus, dans la stratégie d'enregistrement et d'analyse de la substance phonique ainsi que dans le traitement statistique des données nous a permis d'établir une analyse contrastive entre l'espagnol et l'aragonais reposant sur 40 phrases prononcées chacune 10 fois.

Notre objectif étant la mise en évidence de la variabilité de la structure des mots prosodiques en fonction de leur positionnement sur l'axe syntagmatique (incidence de la complexité syntaxique) dans le cadre du schéma général de la phrase, le corpus présente les caractéristiques suivantes :

- Choix d'unités représentatives des principales structures accentuelles admises pour les deux langues (proparoxyton, paroxyton, oxyton).
- À partir de phrases simples (phrases de référence) de type SN + SV ont été opérées plusieurs expansions de chacun des syntagmes par le biais de l'adjonction d'adjectifs et de syntagmes prépositionnels de structure accentuelle variable.
- Afin d'éviter les modifications liées aux caractéristiques intrinsèques et contextuelles, nous avons employé les mêmes unités.

Dans le cadre de cet article, étant donné le peu d'espace disponible, nous nous limitons à présenter les résultats relatifs à quelques structures syntaxiques.

Le corpus a été réalisé, pour l'espagnol, par une locutrice madrilène, et pour l'aragonais, par un locuteur de Bielsa. Concernant le traitement instrumental, les mesures sont issues des tracés de Fo (début, fin et un point intermédiaire pour chacune des voyelles), de l'intensité (point culminant) et du signal de parole pour la durée (Voice Vowel Onset - Voice Vowel Termination). Les données ont fait l'objet d'un traitement statistique permettant le calcul des valeurs moyennes pondérées par les écarts-types. Pour l'aragonais, du fait d'une plus grande instabilité de la durée, nous avons procédé à une analyse de la variance (ANOVA). Pour notre contribution nous présentons uniquement les résultats concernant la varition de Fo.

3. RESULTATS

3.1. Stratégie de l'opposition des modalités

3.1.1. Phrases simples

Les divergences dans les deux langues sont très marquées lorqu'on compare la stratégie de la question totale vs l'assertion.

Notons tout d'abord que le locuteur aragonais utilise dans le cadre de la modalité interrogative un registre moyen légèrement plus étendu que celui de l'espagnol soit 19/4 de ton contre 15/4 de ton. D'autre part, pour

l'interrogation, si l'espagnol se caractérise par une répartition des valeurs de Fo de part et d'autre de la fréquence laryngienne moyenne (figures 1 et 3), l'aragonais en revanche implique une répartition des valeurs située essentiellement au-dessus de la fréquence laryngienne moyenne comme l'illustrent les figures 2 et 4 (partie médiane de la phrase). La ligne mélodique de la phrase interrogative se détache alors très nettement de celle de l'assertion et se situe bien au-dessus du schéma de la phrase affirmative (1,5 ton en moyenne).

Des divergences supplémentaires sont observables. Ainsi, le locuteur espagnol (figures 1 et 3), outre un indice de changement de modalité majeur, repérable en finale absolue (relèvement abrupt de Fo sur la voyelle terminale de l'énoncé autour de 17/4 de ton dans l'interrogation où la montée abrupte atteint les valeurs maximales de la phrase), développe des marques de discrimination modale en amont. A ce propos les schémas mélodiques des voyelles positionnées en fin de SN sont diamétralement opposés (montée pour l'assertion contre une nette décroissance de la fréquence fondamentale pour l'interrogation), ce phénomène se renouvelant au niveau du verbe avec apparition d'un groupe progrédient dans le cadre de l'assertion contre un schéma descendant pour la phrase interrogative. Cette seconde mise en relief dans la phrase affirmative, sans omettre le décrochement entre le SN et le SV (compris entre 4/4 de ton et 7/4 de ton), traduit une plus grande autonomie des unités inférieures (syntagme ou mots prosodiques). La phrase assertive est en effet fracturée en plusieurs endroits de la chaîne (émergence de plusieurs ruptures tonales majeures aboutissant à la constition de 3 intonèmes). En revanche, la modalité interrogative présente un contour global moins heurté qui se caractérise par un continuum décroissant de la fréquence fondamentale à partir de la post-tonique du SN avec chute de Fo qui passe en-dessous de la fréquence laryngienne moyenne, ce niveau étant maintenu jusqu'à la voyelle pénultième.

Pour le locuteur aragonais, nous observons que le changement de modalité ne se traduit plus de la même façon. La séparation entre les deux courbes ne survient plus après la syllabe post-accentuelle mais plus tôt en l'occurrence après la syllabe accentuée du SN (figures 2 et 4). D'autre part, après une montée particulièrement forte dans le cadre de l'interrogation affectant toujours la première post-accentuelle (de 8/4 de ton pour l'unité proparoxytonique, de 9/4 de ton pour la paroxytonique), il y a établissement -toute modalité confondue- d'une zone stationnaire ou palier qui concerne la partie médiane de la phrase. Une seconde rupture apparaît ensuite régulièrement en fin de SV avec une chute abrupte de Fo sur la voyelle accentuée du dernier mot prosodique. La pente très nettement abrupte à partir de la tonique de l'unité accentuelle finale pour l'interrogation, se substitue dans le cadre de l'assertion par une déclinaison en pente douce acompagnée d'un léger relèvement sur la voyelle terminale.

3.2. Les phrases avec expansion

À titre illustratif, nous présentons des expansions en contexte droit, avec syntagme prépositionnel avec le substantif paroxyton, introduites en SN et en SV afin de déterminer les incidences sur la structure accentuelle initiale du mot prosodique vedette.

3.2.1. Expansion en SN

Qu'il s'agisse du représentant espagnol ou aragonais, la configuration générale de la phrase de référence correspondante est globalement conservée.

Pour l'espagnol (figure 5), nous observons néanmoins dans le cadre de la modalité interrogative, une montée plus prononcée sur la syllabe accentuée du premier mot prosodique ainsi que l'émergence d'un petit palier qu'elle forme avec la syllabe subséquente. De plus, l'introduction du syntagme prépositionnel implique une forte chute (de 1,5 ton), amorcée au niveau de la préposition et se prolongeant jusqu'à la frontière non terminale majeure. A signaler également au niveau de la première postonique, une valeur quasi identique à celle de la voyelle terminale (elles constituent les deux pics mélodiques). Concernant l'assertion l'adjonction de l'expansion induit des modifications au niveau du premier mot prosodique qui est alors caractérisé par un contour plus écrasé avec concentration des valeurs autour de la fréquence laryngienne moyenne. Le mot prosodique final accuse lui une chute plus marquée (2 tons contre 1,5 ton dans la phrase simple), avec des valeurs toutes positionnées en dessous de la fréquence laryngienne moyenne.

En aragonais (figure 6), les divergences sont minimes et observables principalement pour l'assertion où la courbe décline sur les syllabes pré-accentuelles et accentuelle du SP, avant de monter sur la post-accentuelle (montée nette et déplacement du sommet de phrase). A remarquer que la chute sur la finale du verbe par rapport à la phrase de référence est beaucoup plus marquée (écart de 1,5 ton).

Signalons en dernier lieu que l'introduction

de l'expansion dans le cadre de l'assertion se traduit par un élargissement du registre.

3.2.2. Expansion en SV

Comme précédemment, l'allure générale des courbes mélodiques demeure. Néanmoins des différences se font jour par rapport aux phrases simples.

Concernant le locuteur espagnol (figure 7), elles sont essentiellement observables dans le cadre de l'interrogation, au niveau de l'augmentation du registre (voyelle finale), de la déclinaison des pentes mélodiques (rupture au niveau de la voyelle initiale accentuée et des segments vocaliques adjacents), du déplacement du sommet mélodique de phrase (sur la post-accentuelle du premier mot prosodique). Quant à l'assertion, le substantif paroxytonique acquière un contour global ascendant avec émergence d'un groupe terminal progrédient indiquant un intonème de continuité et non plus de finalité.

Pour le représentant aragonais (figure 8), l'introduction de l'expansion, sans discrimination de modalité, implique l'émergence de contours plus abrupts au niveau du mot prosodique initial. Pour l'assertion, signalons l'apparition à la frontière non terminale majeure d'une rupture mélodique descendante très forte (de 2 tons). Concernant l'interrogation, nous relevons dans la partie médiane de la phrase (zone stationnaire), un décrochement mélodique décroissant en SV se situant sur la voyelle prétonique du substantif).

4. CONCLUSION

Nous retiendrons que les différences majeures entre les deux systèmes linguistiques apparaissent nettement dans l'opposition modale. D'autre part, l'adjonction des expansions en SN et SV avec changement de frontière syntaxique pour le mot prosodique vedette nous a permis d'illustrer pour chacune des deux langues les modifications éventuelles de la configuration mélodique de base. La complexité des corpus nous permettra d'établir ultérieurement des comparaisons complémentaires en considérant notamment la structure prosodique de l'élément adjectival en fonction de la gradation de la complexité syntaxique.

5. BIBLIOGRAPHIE

Abry C., Benoit C., Boë L. & Sock R. (1985) Un choix d'événements pour l'organisation temporelle du signal de parole, *14ème JEP-GALF*, 133-132

Alcoba S., Le Besnerais M. & Murillo J. (1992) Unité tonale et structure prosodique de l'espagnol, *Revue de Phonétique Appliquée*, n°105, 261-285

Alvar M. (1971) Un problème de langue en contact: la frontière catalano-arargonaise, *Travaux de Linguistique et de Littérature*, IX, n°1, 84

Alvar M. (1973) *Estudios sobre el dialecto aragonés*, I., Zaragoza

Alvar M. (1979) *Atlas linguístico y Etnográfico de Aragón, Navarra y Rioja*, Madrid

Contini M. (1991) Vers une géoprosodie, *Actes du Congrès International de Dialectologie*, 83-109

Delattre P. (1964) *Comparing the Phonetic features of English, German, Spanish and French*, Groos Verlag

Franchon Cabrera C. (1994) *Accent et intonation en castillan: phrases affirmative et interrogative*, Thèse de doctorat ICP, Université Stendhal, Grenoble

Navarro Tomas T. (1948) *Manual de pronunciacion española*, Hispanic Institute

Quilis A. (1979) Fonction linguistique de l'intonation, *Travaux de l'Institut de Phonétique de Strasbourg*, n°11, 79-108

Quilis A. (1993) *Tratado de fonología y fonética españolas*, Gredos

Rhardisse A. (1994) *Accent et intonation du parler de Bielsa (Aragon-Espagne)*, Thèse de doctorat , Université Stendhal, Centre de Dialectologie, Grenoble

Rossi M. (1985) L'intonation et l'organisation de l'énoncé, *Phonetica*, n°42, 135-153

Stockwell R.P., Bowem J. D. & Silva-Fuenzalida I. (1956) Spanish Juncture and Intonation, *Langage*, n°32, 641-665

Espagnol / Tu chiquita mira al pájaro

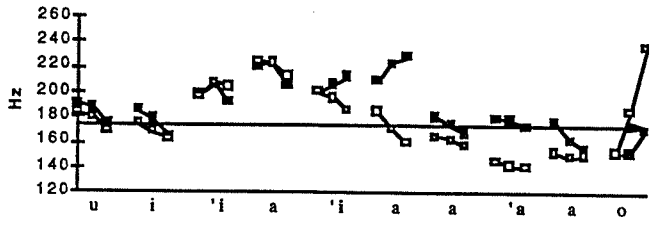


Figure 1: Phrase simple

Aragonais / El misache mira al pájaro

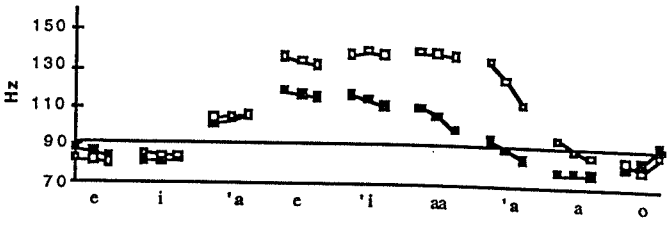


Figure 2 : Phrase simple

Espagnol / Tu pájaro mira a la chiquita

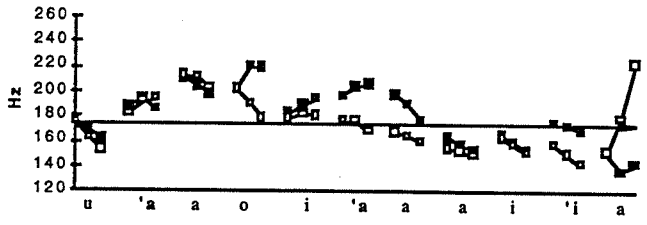


Figure 3: Phrase simple

Aragonais / El pájaro mira al misache

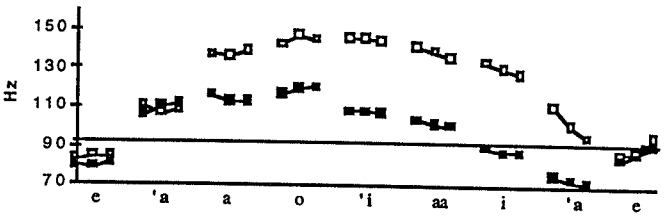


Figure 4: Phrase simple

Espagnol / Tu chiquita de Toledo mira al pájaro

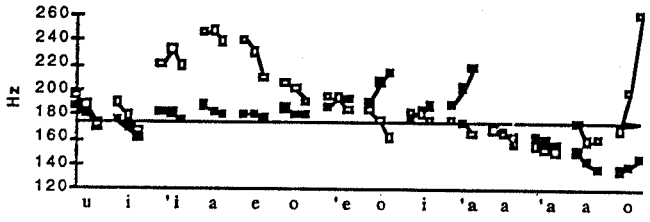


Figure 5: Expansion en SN

Aragonais / El misache de Toledo mira al pájaro

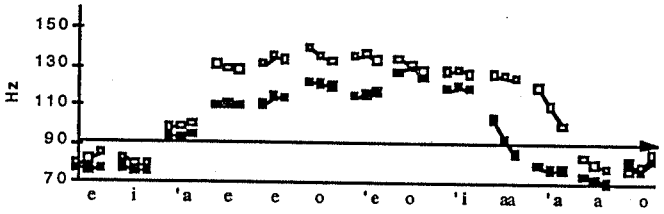


Figure 6: Expansion en SN

Espagnol / Tu pájaro mira a la chiquita de Toledo

Figure 7: Expansion en SV

Aragonais / El pájaro mira al misache de Toledo

Figure 8: Expansion en SV

■ Affirmation
❑ Interrogation

■ Affirmation
❑ Interrogation

ETUDE DU RYTHME DE L'ANGLAIS ET DU FRANÇAIS : ANALYSE D'UNE CHANSON EN DEUX LANGUES

Bertrand LAURET
Institut de Linguistique et de Phonétique Générales et Appliquées, URA 1027 / DFLE
Université Paris III
19, rue des Beranardins 75005 - Paris

ABSTRACT
Following the rhythmic frames of a "rap" song, the prosodic structures (rhythm and stress) of the English and French lyrics of this song were compared. For most utterances (93% in English and 84% in French), the preservation of the prosodic caracteristics of each language was observed : the stressed syllable of the final lexical word of the sense groupe in English and the last syllable of the sense group in French are, at least, under the rhythmic pulsation. Some rhythmic effects non-coinciding with the prosodic structures (compensated by use of duration for both languages) were described as creating the style of the song.

1. INTRODUCTION

On a souvent décrit ou contesté l'opposition langue accentuelle ("stress-timed language") et langue syllabique ("syllable-timed language") entre l'anglais et le français [PIKE, 1945 ; ABERCROMBIE, 1964 ; WENK et WIOLAND, 1982 ; DAUER, 1983 ; FLETCHER, 1991 ; FANT, 1991]. D'autres terminologies ont été proposées pour le français : "trailed timed language" (l'élément saillant en français serait situé à la frontière droite du groupe de sens et non pas, comme en anglais, à la frontière gauche du pied initial) [FLETCHER, 1991] ou "isochronie entre les coupes" [FRAISSE, 1956] ou "boundary language" [VAISSIÈRE, 1991]. Le rythme serait essentiellement dû à une égalité perçue de la durée des syllabes en français et à une tendance à l'isochronie des syllabes accentuées en anglais. Quelles sont les recherches ayant abouti à ces conclusions ? Y a-t-il des cadres d'observation propices à la comparaison du (des) rythme(s) de deux langues?

2. RYTHME ET ACCENT EN FRANÇAIS ET EN ANGLAIS

On présente souvent le français comme une langue à accent de mot intrinsèque (de durée) porté sur toute la dernière syllabe de tout mot lexical (les mots grammaticaux - articles, auxiliaires, n'ont pas de dernière syllabe accentuable). En parole spontanée, ces accents de mots ne sont pas ou en partie réalisés. C'est alors la dernière syllabe d'un groupe de mots, constituant un groupe de sens (accent de groupe), qui est allongée (dans un rapport de 2,1 pour les CV et 1,9 pour les CVC, [WENK et WIOLAND, 1982]). Suivant la vitesse d'élocution et le locuteur, l'énoncé est découpé en un nombre différent d'unités de sens [DELATTRE, 1965 ; WIOLAND, 1991].

La position de l'accent (intensité) dans les mots en français n'est pas distinctive. Mais une proéminence sur la première syllabe comme une variante (régionale, stylistique) a été décrite [GRAMMONT, 1914 ; LUCCI, 1983]. La plupart des phonéticiens négligent la proéminence sur les syllabes initiales. Pourtant, [VAISSIÈRE, 1974] décrit une intensité initiale chez des locuteurs non professionnels sans accent régional et [FONAGY, 1980] considère l'intensité initiale comme une marque de changement en cours du système prosodique du français. Le français se caractériserait donc par un surcroît de durée sur la dernière syllabe du groupe de sens, et par un surcroît d'intensité (pas toujours réalisé) sur la syllabe initiale du groupe.

En anglais, les syllabes sont nettement inégales en durée et en intensité. L'accent en anglais a une fonction distinctive (/'ɒb-dʒɪkt/ "objet" ; /əb-'dʒekt/ "faire une objection"). La position de l'accent est inégalement répartie : 74% des mots bisyllabiques ont un accent initial [DELATTRE, 1965]. La prédominance de l'accent initial influence fortement les stratégies de perception des mots anglais. Les anglophones percevraient plus rapidement les mots dont l'accent est placé sur la première syllabe que les mots dont l'accent est placé sur une autre syllabe, en segmentant donc le continuum sonore avant les syllabes accentuées. L'unité rythmique est généralement décrite comme un groupe de syllabes dont une seule est forte ("strong" vs "weak"). Un rapport de durée de 1 à 1,8 entre les syllabes non accentuées et les syllabes accentuées a été décrit par Wallin [cité par FRAISSE, 1956]. Si les langues syllabiques sont perçues comme présentant des syllabes de même durée, les langues accentuées seraient perçues comme possédant des "pieds" isochrones. [SCOTT, ISARD et BOYSSON-BARDIES, 1985] ont montré que l'isochronie ne semble pas caractériser les langues accentuées. N'ayant pu mettre en évidence la rigueur du concept d'isochronie,

[FLETCHER, 1991] aboutit au concept d'isochronie faible. Pour [FANT, 1991], à la suite d'autres auteurs, le concept de langue accentuelle n'est pas dû à une isochronie physique mais à la relative proéminence syllabique des accents, alors qu'une langue syllabique comme le français doit sa régularité à la prédominance de syllabes CV et au faible degré d'accent induit par l'allongement segmental.

Le principe de base du rythme dans la parole est probablement indépendant de la langue étudiée : le rythme est produit par la répétition de schémas composés d'une succession d'événements. On peut donc penser que le rythme ne se réduit pas à la seule perception d'un accent (quelque soit sa nature) sur une syllabe, mais que les composantes du rythme peuvent être hiérarchisées.

Pour résumer, on peut dire qu'en anglais l'accentuation ("stressed syllable") est dominante, alors que l'allongement syllabique est "latent". En français, c'est l'organisation temporelle qui est dominante mais l'accentuation initiale est aussi intrinsèquement présente.

3. RYTHME, POESIE ET MUSIQUE

Y a-t-il des cadres d'observation propices à la comparaison du (des) rythme(s) de deux langues?

Pour [FRAISSE, 1956], les structures rythmiques de la poésie ont été basées sur le principe de différenciation des longues et des brèves (quand c'est possible, longues et brèves présentant un rapport de durée de 2), ou sur le retour périodique d'accents ou de coupes. La périodicité dominante est celle de l'isochronisme entre les accents ou les coupes ce qui correspond aux deux rythmes de base : groupement et organisation de durées ; groupement de "moments privilégiés" (accents en général) avec lequel on peut trouver des groupements de durées.

L'étude de la versification confronte les structures métriques (nombre et suite de pieds) et les structures prosodiques. La non-coïncidence entre une catégorie métrique (vers, pied) et la catégorie prosodique correspondante (énoncé, mot prosodique, syllabe) est sujette à des contraintes qu'on cherche à établir. Quand la contrainte est violée, on peut penser (i) que la structure prosodique de la phrase est incorrecte, (ii) que la contrainte établie est incorrecte, (iii) que la contrainte a été délibérément violée par l'auteur [VERLUYTEN, 1989]. De la même manière pour l'anglais, [KIPARSKY, 1975] effectue pour chaque vers une comparaison entre structure métrique et structure prosodique, et observe les concordances et discordances. Les discordances (la complexité) caractérise la limite de non-métricité que l'auteur s'impose.

Cette méthode peut être utilisée en musique, si l'on compare une structure rythmique musicale et une structure prosodique.

4. ANALYSE D'UNE CHANSON "RAP"

4.1. Objectifs

On a souvent vanté les capacités musicales des langues accentuelles (comme l'italien) par rapport aux langues syllabiques (comme le français). Ainsi, Rousseau conclut sa *Lettre sur la musique française* (1753) : "D'où j'en conclus que les Français n'ont point de musique et n'en peuvent avoir, ou que si jamais ils en ont une, ce sera tant pis pour eux".

Nous avons choisi d'observer les rythmes de l'anglais et du français confrontés aux cadres rythmiques de la musique [HEFFNER, 1950 ; FRAISSE, 1956 ; VAISSIERE, 1991]. Pour que le cadre rythmique musical soit le plus similaire possible pour les deux langues, nous avons analysé un "rap", genre de chanson qui consiste à syllaber le texte avec une voix relativement monocorde sur une pulsation régulière réalisée par des percussions. Ce rap présente la particularité d'être chanté en deux langues, alternativement en anglais et en français par des locuteurs natifs dans la même chanson [URBAN SPECIES, 1994]. Une comparaison des rythmes musicaux employés par les deux langues sera effectuée.

4.2. Corpus

La chanson présente alternativement des passages en anglais et en français (il ne s'agit pas de traductions), au total un texte de 287 mots en anglais, et 245 mots en français (une pulsation musicale tous les 2,08 mots en moyenne en anglais, 2,15 mots en moyenne en français). Suivant la partition de percussion, la mesure est binaire, présentant 4 noires (N) (mesure à 4/4). La durée la plus utilisée est la double croche (dC). Une notation présentant la transcription phonétique des syllabes suivant le rythme (en gras, les syllabes sous la pulsation musicale) a été faite. en anglais :

N				N			
dC1	dC2	dC3	dC4	dC1	dC2	dC3	dC4
ˈlɪ-	sn	tu:	ðə	**ˈrɪ-**	ðm		

Listen to the rhythm

N				N			
dC1	dC2	dC3	dC4	dC1	dC2	dC3	dC4
ˈlɪ-	sn	tu:	ðə	**raɪm**			

Listen to the rhyme

en français :

N				N			
dC1	dC2	dC3	dC4	dC1	dC2	dC3	dC4
ki	ɛ	la-	ni-	**mal**			lwa-

Qui est l'animal

N				N			
dC1	dC2	dC3	dC4	dC1	dC2	dC3	dC4
zo	u	lə	ti-	**rœr**			

l'oiseau ou le tireur

4.3. Résultats

4.3.1. Les tendances générales

En anglais, on trouve sous la pulsation des monosyllabes (lexicaux le plus souvent, WORLD, RHYME, HEART) ; des mots lexicaux de plus d'une syllabe (2 syllabes, 'RYTHM, 'SOLVING, 3 syllabes, E'VOLVING, 'BEGGAR-MAN, 6 syllabes, RESPONSA'BILITIES). On observe que la syllabe accentuée est placée sous la pulsation dans 93% des polysyllabes du texte (67/72).

	puls.		
	'sɒl-	vɪŋ	SOLVING
	'beg-	ɜ-mæn	BEGGARMAN
	'pe-	sɪ-mɪst	PESSIMIST
kən-	'trəʊl		CONTROL
ɪ-	'vɒl-	vɪŋ	EVOLVING
rɪs-pɒns-ɪ-	'bɪ-	lɪ-tɪz	RESPONSIBILITIES

En français, on trouve sous la pulsation des monosyllabes (lexicaux le plus souvent BALLE, CIEL) ; des mots lexicaux de plus d'une syllabe (2 syllabes, MALADE, DOMAINE, 3 syllabes, ANTIDOTE, BAVARDAGE). On observe que la syllabe finale est placée sous la pulsation dans 84% des polysyllabes du texte (46/55).

		puls.	
	ma-	lad	MALADE
	sy-	syr	SUSSURE
lã-	ti	dɔt	L'ANTIDOTE
ba-	var-	daʒ	BAVARDAGE
la-	ni-	mal	L'ANIMAL
kɔ̃-	sã-	sys	CONSENSUS

Ces placements des polysyllabiques anglais et français sur le rythme musical illustrent l'importance accordée ici à la syllabe accentuée en anglais et à la dernière syllabe en français confirmant empiriquement les principales tendances décrites pour la parole dans les deux langues (proéminence de la syllabe accentuée en anglais, proéminence de la dernière syllabe en français). La régularité syllabique et la syllabation en français, parfois travaillées dans les manuels de phonétique française à l'usage des anglophones, trouvent ici une illustration exploitable en classe de langues (où les exercices de rythme sont rares souvent par défaut de support rythmique "musical").

4. 3.2. Les autres configurations

Certaines configurations semblent n'exister qu'à cause des contraintes de la mesure musicale, alors que d'autres configurations (RARE) non contraintes caractérisent le style de la chanson.

4.3.2.1. Les contextes contraignants (pour préserver l'accent de groupe)

Les mots grammaticaux sous la pulsation en français (EST-CE BIEN, QUI (interrogatif) J'ÉTAIS) sont naturellement accentuables [VERLUYTEN, 1989].

Les mots grammaticaux sous la pulsation en anglais (OF, HIS, WHO) préservent la pulsation suivante sur une syllabe accentuée.

En anglais, on observe des dissyllabiques hors pulsation ('MOTHER), pour permettre à un autre mot lexical, en position finale du groupe de sens, d'être sous l'accent.

N				N			
dC1	dC2	dC3	dC4	dC1	dC2	dC3	dC4
əv	ðə	'mʌ-	ðə	**ɜːθ**			

of the mother earth

De la même façon, on trouve, en français, des dissyllabiques dont la syllabe initiale est accentuée (SYSTÈME, SOLDAT) pour conserver la pulsation sur la dernière syllabe du groupe de sens.

	N					N			
dC4	dC1	dC2	dC3	dC4		dC1	dC2	dC3	dC4
lə	**sis-**	tɛ-	mɛ	ma-	**lad**				

Le système est malade

(...)

	N					N			
dC4	dC1	dC2	dC3	dC4		dC1	dC2	dC3	dC4
dœ̃-	**sɔl-**	da	ɛ̃-	ko-	**ny**				

d'un soldat inconnu

4.3.2.2. Les contextes non-contraignants ou les jeux de rythme propres au style "rap"

On observe en anglais des polysyllabiques dont c'est la syllabe inaccentuée qui est située sous la pulsation (ex : 'JUSTICE, 'SCAPEGOAT) et ce, sans contrainte du contexte. Cet effet est compensé par la durée de la syllabe accentuée (2dC). L'environnement (CON'TROL, 'GREATER) est préservé.

N				N			
dC1	dC2	dC3	dC4	dC1	dC2	dC3	dC4
let	ɪt	teɪk	kən-	**'trəʊl**	əv	'dʒʌs-	

Let it take control of

N			
dC1	dC2	dC3	dC4
tɪs			

justice

(...)

N				N			
dC1	dC2	dC3	dC4	dC1	dC2	dC3	dC4
'greɪ-tər		'skeɪp-		**gəʊt**			

greaterscapegoat

En français, on trouve également des mots dont la dernière syllabe ne tombe pas sur la pulsation (CONCRET, ABSTRAIT) sans contrainte du contexte. Cet effet est, comme en anglais, compensé par la durée (2dC).

N N
dC1 dC2 dC3 dC4 dC1 dC2 dC3 dC4
- ə dy kɔ̃- - krɛ
(ils veulent) du concret
(...)
N N
dC1 dC2 dC3 dC4 dC1 dC2 dC3 dC4
si- lɛ- taps- - trɛ
(même) s'il est abstrait.

Ces syncopes non contraintes (avec allongement) présentes en anglais comme en français semblent caractéristiques du style "rap". Elles ne coïncident pas avec la prosodie naturelle de la langue (les résultats d'un test de placement de l'accent à partir d'extraits du texte seul auprès de 25 adultes français natifs confirme que la pénultième d'un groupe de sens n'est pas "naturellement" accentuable).

4.3.3. Rythme musical

Les structures rythmiques musicales pour chaque énoncé-groupe de sens (séparé par deux pauses) ont été transcrites. L'accent anglais frappant le plus souvent la première syllabe du mot et en particulier des très majoritaires dissyllabiques (74%, [DELATTRE, 1965]), on constate une tendance à un rythme qui commence sur le temps. Le rythme ci-dessous (Fig.1) et ses variantes (remplaçant la noire, deux croches, deux, trois ou quatre double croches), est utilisé dans 71% des cas (48/68). On relève peu d'énoncés commençant par une levée.

Fig.1- Rythme musical le plus utilisé (71%) par le texte anglais

En français, la pulsation est perçue sur la dernière syllabe. Le rythme le plus utilisé (Fig.2) commence une phrase sur la levée (anacrouse) dans 72% des cas (40/55).

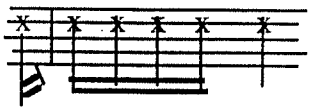


Fig.2 - Rythme musical le plus utilisé (72%) par le texte français

5. CONCLUSION

L'analyse du texte d'une chanson "rap" en anglais et en français a montré que, lorsque la répartition des syllabes des deux langues est réalisée sur une isochronie musicale (rythme régulier de la pulsation), l'anglais met très majoritairement en valeur la syllabe accentuée alors que le français met très majoritairement en valeur la dernière syllabe de mots lexicaux. En cas de conflit entre deux mots lexicaux, les deux langues privilégient la syllabe accentuée du dernier mot du groupe de sens. Seuls quelques jeux de rythme (toujours différents et représentant 10% des phrases) caractérisent ici le style "rap" de la chanson. Ce style compense la non-coïncidence des structures musicales et prosodiques par des allongements de durée de même type dans les deux langues.

6. BIBLIOGRAPHIE

ABERCROMBIE, D., *Elements of General Phonetics*, Edinburgh University Press, 1967.

DAUER, R.M., Stress-timing and syllabic-timing reanalyzed, *Journal of Phonetics*, 11, 1983, pp. 51-62.

DELATTRE, Pierre, *Comparing the Phonetic Features of English, French, German and Spanish*, Julius Groos Verlag, Heidelberg, 1965.

FANT, Gunnar, KRUCKENBERG, Anita, NORD, Lennart, Durational correlates of stress in Swedish, French and English, *Journal of Phonetics*, 19, 1991, pp. 351-365.

FLETCHER, Janet, Rythm and final lengthening in French, *Journal of Phonetics*, 19, 1991, pp. 193-212.

FONAGY, Ivan, L'accent français : accent probabilitaire, *Studia Phonetica* 15, Fonagy & Léon éd., 1980.

FRAISSE, Paul, *Les structures rythmiques*, Louvain, PULouvain, 1956

GRAMMONT, G., *Traité pratique de prononciation française*, Paris, 1914

HEFFNER, R-M. S., *General Phonetics*, Madison, The University of Winconsin Press, 1964.

KIPARSKY, Paul, Stress, Syntax and Meter, *Language*, vol.51, 3, 1975, pp.576-616.

LUCCI, Vincent, *Etude phonétique du français contemporain à travers la variation situationnelle*, Publications des Universités des Langues et ds Lettres, Grenoble, 1983.

PIKE, K.L., *The intonation of American English*, Ann Arbor, University of Michigan Press, 1945.

SCOTT, Donia R., ISARD, S.D., de BOYSSON-BARDIES, Bénédicte, Perceptual isochrony in English and in French, *Journal of Phonetics*, 13, 1985, pp. 155-162.

URBAN SPECIES : Listen (Edit, Alternative Mix), Phonogram Ltd, London, 1994.

VAISSIERE, Jacqueline, On French Prosody, *Quaterly Progress Report*, MIT, Res. Lab. of Electronics, n°114, Juin 1974, 212-223.

VAISSIERE, Jacqueline, Rythm, accentuation and final lengthening in French, *Music, Language, Speech and Brain*, 1991.

VERLUYTEN, Paul, L'analyse de l'alexandrin : mètre ou rythme ? in *Le souci des apparences*, Dominicy éd., Editions de l'Université de Bruxelles, 1989, pp. 31-74.

WENK, Brian J. , WIOLAND, François, Is French really syllabe-timed?, *Journal of Phonetics*, 1982, 10, p. 193-216.

WIOLAND, François, *Prononcer les sons du français*, Hachette, 1991.

PROÉMINENCE DE SYLLABES ET DE FRONTIÈRES EN SYNTHÈSE VOCALE DE L'ALLEMAND

Barbara HEUFT, Thomas PORTELE
Institut für Kommunikationsforschung und Phonetik (IKP)
Universität Bonn, Allemagne

ABSTRACT

The structure of a synthesis system is described that uses prominence as a central parameter. A definition of prominence suitable for this application is given. For the empirical foundation the reliability of prominence ratings by human listeners is assessed. These ratings were compared with acoustic data on F_0 and duration. A linear relationship between ratings and parameter values was found. The application of prominence to some linguistic phenomena is demonstrated. The results indicate the validity of the prominence based approach as an interface between linguistics and acoustics.

1. MOTIVATION

La plupart des systèmes de synthèse produisent de la parole en utilisant comme *"input"* uniquement l'orthographie. Ainsi, une certaine partie d'informations n'est pas accessible, p.ex. des informations sur la pragmatique, la structure du discours etc. Toutefois, dans les systèmes d'information qui contiennent un module de génération de texte, une certaine quantité de ces informations est présente. Afin de pouvoir aisément traiter ces informations dans un système de synthèse, nous introduisons un paramètre intermédiaire, que nous allons appeler la *proéminence*.

2. PROÉMINENCE

La proéminence est tout d'abord un paramètre perceptif: *"Prominence is the property by which linguistic units are perceived as outstanding from their environment."* (Terken, 1991). C'est ensuite un paramètre graduel "... *accent prominence is relative, and a matter of degree: some accents are more prominent than others.*" (Ladd et al., 1994).

Dans notre application, nous donnons à la proéminence la définition suivante (Portele & Heuft, 1996):

La proéminence est un paramètre quantitatif attribué à une syllabe ou à une frontière. C'est un paramètre du système qui correspond à la proéminence perçue par un auditeur humain. Quand un rangement de syllabes et frontières, dont à chacune est attribuée une valeur de proéminence, est synthétisée, un auditeur devrait percevoir les relations de proéminence parmi les syllabes et frontières synthétisées d'une façon qui reflet les valeurs numériques du paramètre du système.

C'est surtout le traitement des paramètres prosodiques qui va permettre la perception appropriée de la proéminence. Figure 1 montre la structure d'un système de synthèse vocale basée sur la proéminence. Dans un tel système, chaque syllabe et chaque frontière n'a que deux propriétés: Le contenu (les phonèmes) pour les syllabes, le type (progrédient ou non-progrédient) pour les frontières et la proéminence. La supposition principale c'est que cette déscription est suffisante. Afin de prouver cette hypothèse, il faudra répondre aux questions suivantes:

Est-ce que des auditeurs peuvent juger la

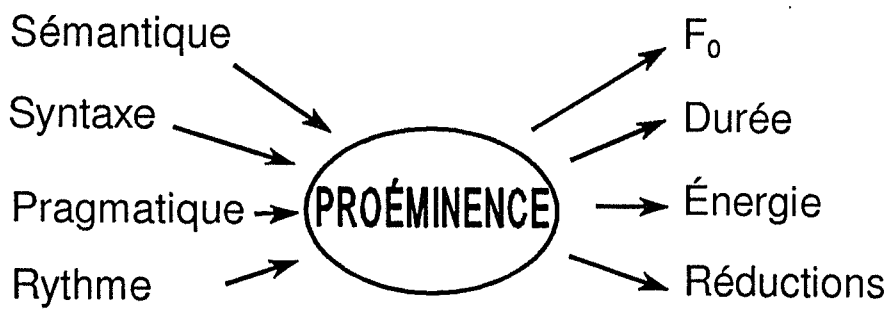


Fig. 1 : Structure d'un système de synthèse vocale basé sur la proéminence

proéminence, dans la gamme choisie dans notre définition, avec une concordance suffisante?

Quel est la relation entre la proéminence perçue et les propriétés acoustiques du signal?

En plus, la transformation des valeurs de proéminence en paramètres prosodiques doit être implémentée et évaluée.

3. JUGEMENTS D'AUDITEURS

3. 1 Syllabes

Dans la première expérience, la proéminence de plus que 11.000 syllabes a été jugée par 3 auditeurs. Les jugements ont été faits sur une échelle graphique s'étendant de 0 à 31 (Fant & Kruckenberg, 1981). Les sujets pouvaient écouter aux stimulis aussi souvent que nécessaire. Le matériel étiqueté comprenait des énoncés isolées, des paires de question et réponse et trois textes. Il était lu par trois locuteurs (1m2f).

Figure 2 montre que la corrélation entre les jugements des 3 sujets est assez haute, environ 0,8 pour tous les locuteurs et tous les sujets.

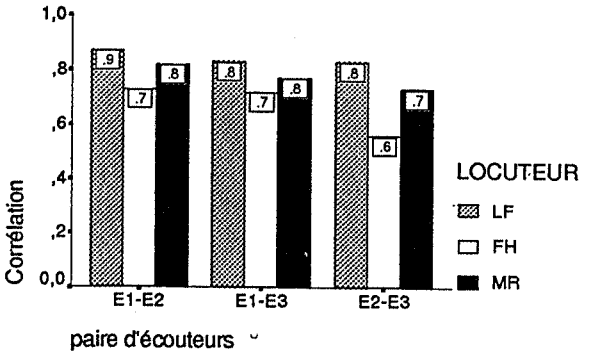


Fig. 2: Coéfficients de corrélation entre les trois sujets A1-A3 jugeant la proéminence de syllabes

3.2 Frontières

Une méthode pareille à la première expérience a été employée (De Pijper & Sanderman, 1994). Cette fois-ci, les trois textes du corpus ont été utilisés. Trois sujets étaient demandés de juger la proéminence de la frontière derrière chaque mot sur une échelle de 0 à 9. Il s'agisseait de 1336 frontières potentielles. Comme pour la proéminence des syllabes, les sujets montraient une grande concordance dans leurs jugements (voir fig.3). Comme réponse à notre première question, on peut donc conclure que la capacité de discrimination des auditeurs est assez grande et que leurs jugements sont assez concordants pour la gamme que nous avions choisie.

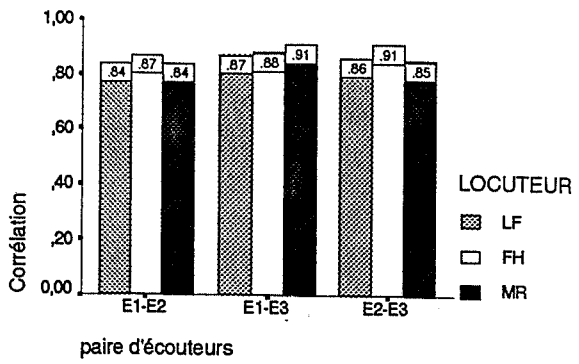


Fig. 3: Coéfficients de corrélation entre les trois sujets A1-A3 jugeant la proéminence des frontières

4. LA RELATION PERCEPTION-PARAMÈTRES PROSODIQUES

Les valeurs de proéminence ainsi déterminées ont été mises en relation avec les paramètres prosodiques.

4.1 Syllabes

Il y a d'abord une relation quasiment linéaire entre durée des syllabes et proéminence perçue. Figure 4 montre clairement le caractère graduel de la proéminence.

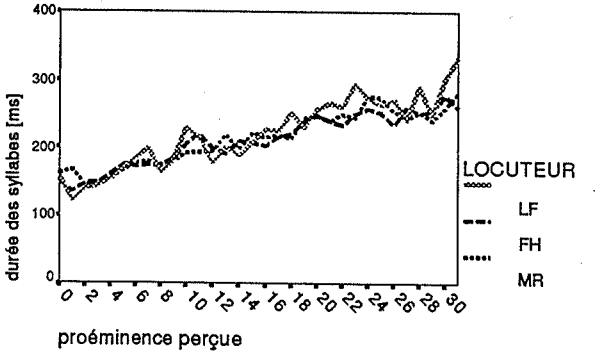


Fig. 4: Relation entre le médian de la proéminence étiqueté et la durée des syllabes

Le deuxième paramètre étudié était la fréquence fondamentale. Comme propriété binaire d'une syllabe il y a la présence d'un pic F_0. Le médian de la proéminence perçue est 15 points plus hauts pour les syllabes asscociés avec un pic F_0 que pour les autres syllabes (voir fig. 5). 68% des syllabes avec un sommet de F_0 ont été attribués en moyenne une valeur de proéminence plus grand que 14.

Mais non pas seulement la présence d'un pic F_0, mais aussi sa hauteur influe la perception de la proéminence. La figure 6 montre la relation entre la proéminence perçue des syllabes associès avec un pic F_0 et la hauteur des ces pics.

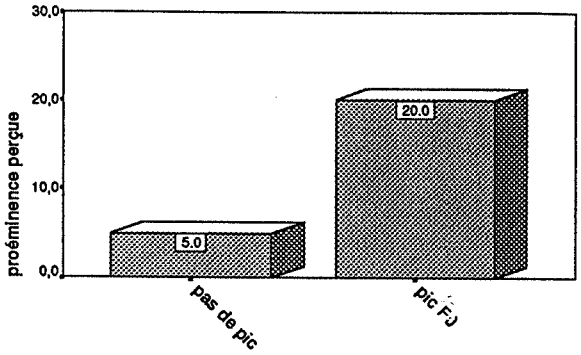


Fig. 5: Influence de la présence d'un pic de F_0 sur la proéminence perceptive

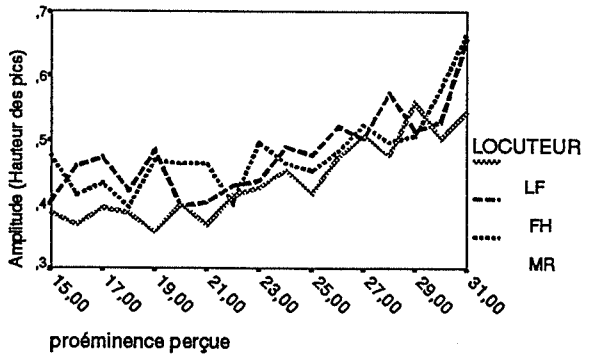


Fig. 6: Relation entre la proéminence étiquetée aux syllabes et la hauteur des pics F_0. Les résultats pour les valeurs au dessous de 15 sont incertains à cause du nombre limité d'occurences.

A part la hauteur des sommets, c'est aussi leur position relative au début de la voyelle de la syllabe accentuée qui exerce une influence sur la perception. Les pics qui se situent plus terminal ou bien derrière la syllabe ont tendance à évoquer une proéminence plus forte que les pics avant ou au début de la voyelle (voir aussi les résultats de Kohler, 1991). Toutefois, cette relation ne peut être constatée que pour quelques positions spécifiques des sommets de F_0 dans

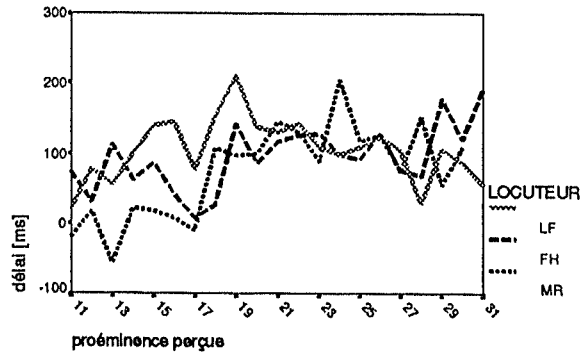


Fig. 7: Relation entre la proéminence des syllabes et le placement des pics F_0 pour le premier sommet dans la phrase (Délai négatif: le pic se produit avant le d'but de la voyelle. Délai positif: le pic se produit après le début de la voyelle).

l'énoncé, p.ex. le premier sommet (voir fig.7). En plus, il semble que des différences inter-locuteurs sont plus nettes que pour les autres paramètres acoustiques.

4.2 Frontières

Ainsi que pour la proéminence des syllabes, nous avons cherché les équivalents acoustiques de la proéminence perceptive des frontières prosodiques (Heuft, 1996). Comme indicateur le plus fort d'une frontière prosodique nous avons trouvé la présence d'une pause. On peut constater une relation directe entre la durée des pauses et la proéminence (voir fig. 8).

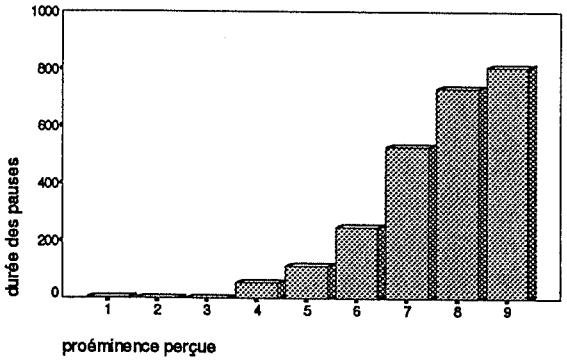


Fig. 8 Relation entre la proéminence étiquété aux frontières (médian) et la durée des pauses

Pour les frontières faibles qui ne sont pas marquées avec une pause, nous avons trouvé une relation entre la durée de la dernière syllabe dans la phrase et la proéminence (voir fig. 9).

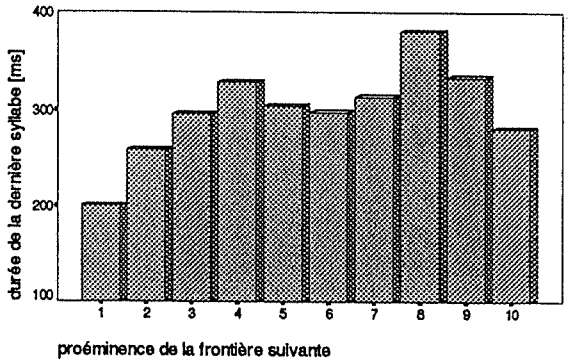


Fig. 9: Relation entre la proéminence étiquetée aux frontières et la durée de la dernière syllabe de la phrase précédente

Il y a une légère corrélation négative entre la hauteur du dernier sommet de la phrase prosodique et la proéminence de la frontière (pour les phrases terminales; $\rho \approx -0{,}32$). Pour les phrases avec une montée de F_0, nous n'avons pas trouvé d'indicateur mélodique; ceci est peut-être dû au

fait que le corpus employé contenait des phrases progrédientes mais non pas de questions.

5. CONCLUSIONS POUR LA SYNTHÈSE

Les résultats décrits ci-dessus impliquent qu'il devrait être possible de modeler la proéminence équivalente à la perception. La composante prosodique d'un système de synthèse destiné à modeler ce phénomène doit pouvoir traiter la proéminence comme paramètre graduel. Le système dévéloppé à Bonn, dont la structure est présentée plus haut cherche à prédire les valeurs des différents paramètres prosodiques dans un seul pas, car ils sont les dépendants de la même variable. Pour les valeurs de F_0, on cherche à prédire surtout la place et la hauteur des sommets de F_0 (voir Heuft et al., 1995; et Heuft, 1995 pour une déscription détaillée).

Il existent deux approches parallèles à la transformation des degrées de proéminence en paramètres prosodiques: Une approche statistique avec un réseau de neurones et une approche qui cherche à intégrer les résultats présentés ici dans des règles.

6. UNE APPLICATION

Une application possible c'est la modélisation adéquate des accents focals. Pour faire une évaluation préliminaire des capacités du système, nous avons effectué une expérience: 5 énoncés étaient synthétisés en 4 versions (5 versions pour une phrase), dans chaque version l'accent focal était placé sur un mot différent. La génération de la prosodie a été effectuée par règles. Les énoncés étaient interprétables comme réponses à des questions qui demandaient les mots focalisés. (p.ex. la question correspondante à la phrase: *Ich bin mit Barbara nach Darmstadt gefahren.* [C'est moi qui est parti à Darmstadt avec Barbara] serait: *Wer ist mit Barbara nach Darmstadt gefahren*? [Qui est parti à Darmstadt avec Barbara?]). 12 sujets étaient demandés de choisir la question correspondante à la phrase qu'ils venaient d'écouter. 4 questions possibles étaient présentées sur une feuille. En tout, il s'agissait de 21 stimulis.

Figure 10 montre les résultats. La plupart des sujets étaient capables d'assigner les questions correctement. Ces résultats sont certainement très positifs, quoiqu'ils ne donnent aucune évidence sur la naturalité des énoncés. Afin de pouvoir interpréter correctement ces résultats, il reste à effectuer la même experience avec un locuteur humain.

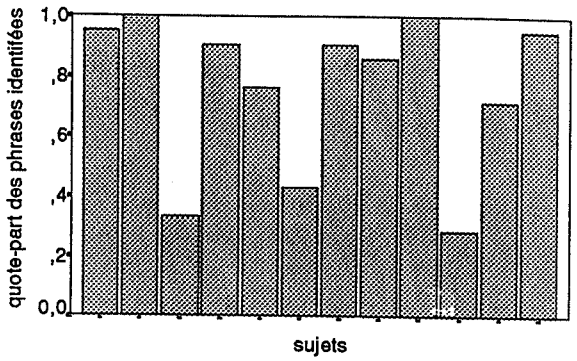


Fig. 10: Identifications correctes des phrases avec différentes localisations de l'accent focal

7. DISCUSSION

Les résultats sont, quoique préliminaires, assez prometteurs. La proéminence est certainement un paramètre graduel dont les grades sont reflets dans les paramètres prosodiques. Pour évaluer la synthèse, il faudra répéter les mêmes expériences avec la parole synthétisée. Il faudra voir si les valeurs des proéminence du système et les valeurs perceptives seront égales.

Comme plus grand problème de l'avenir reste la prédiction des niveaux de proéminence à partir des informations linguistiques.

BIBLIOGRAPHIE

Fant, G.; Kruckenberg, A.(1989): Preliminairies to the study of Swedish prose reading and reading style. *STL-QPSR* 2/1989, 1-53

Heuft, B.; Portele, T.; Höfer, F.; Krämer, J.; Meyer, H.; Sonntag, G. (1995): Parametric description of F_0-contours in a prosodic database. *Proc. of the XIIIth ICPhS*, Stockholm, p.378-381, vol.2

Heuft, B. (1996): Predicting F_0-peak position. Proc. of Forum Acusticum'96, Antwerpen

Heuft, B.; Rauth, M.; Höfer, F. (1996): Prominenz von prosodischen Grenzen. DAGA'96, Bonn

Terken, J. (1991): Fundamental frequency and perceived prominence of accented syllables. *J.Acoust. Soc. Am.* 89 (4), 1768-1776

Kohler, K.J. (1991):Prosody in speech synthesis: The interplay between basic research and TTS application. *Journal of Phonetics* 23, p. 429-451

Ladd, D. R.; Verhoeven, J.; Jacobs, K. (1994): Influence of adjacent pitch accents on each other's perceived prominence: two contradictory effects. *Journal of Phonetics* 22, 87-99

De Pijper, J.R.; Sanderman, A. (1994): On the perceptual strength of prosodic boundaries and its relation to suprasegmental cues. *J.Acoust. Soc. Am.* 96 (4), 2037-2047

Portele, T.; Heuft, B. (1995): Towards a prominence based synthesis system. Proc. 2ns Speak!-workshop

REMERCIEMENT

Ces travaux ont été soutenus par le ministère de la recherche allemand (BMBF) dans le cadre du projet Verbmobil; contrat Nº 01IV101N0.

PROGRAMMATION DE LA PRODUCTION ET ANTICIPATION DE L'IDENTIFICATION DES FORMES PROSODIQUES ETUDE DEVELOPPEMENTALE

Juliette CLEMENT & Claire GERARD

Laboratoire de Psychologie Expérimentale - 28, Rue Serpente - 75006 Paris
Tél.: 40 51 98 65 - Fax : 40 51 70 85

ABSTRACT

Groups of 5, 7 and 9 year-old children were compared to adults in two experiments conducted to test the existence of a "prosodic lexicon". The first experiment investigated prosodic regulations in the unfolding of the discourse by studying the production of illocutionary forms. The second experiment examined the identification of illocutionary forms within a gating procedure. We observed that only 9 year-old children were similar to adults as regards the speech local programmation at the beginning of the sentence. At this age, however, there was no evidence that children could really anticipate the local recognition of illocutionary forms, even though they obtained the same percentage of global correct responses as adults. This suggests the existence of a difference between production and perception of prosody in the development of speech comprehension mechanisms.

1. INTRODUCTION

L'action de parler ("speech act") englobe le fait qu'un locuteur produit des phrases dans le but de réaliser une certaine "intention communicative", qui n'est pas exprimée par le contenu lexical de l'énoncé : la **force illocutoire** est le message sous-jacent à l'énoncé, se distingue de la fonction locutoire, message transmis par les mots mêmes et de la fonction perlocutoire, conséquence du message (Caron, 1983, 1989). A l'interprétation sémantique s'ajoute ainsi un second processus, l'interprétation pragmatique, consistant à rechercher les intentions du locuteur, exprimées par des schémas intonatifs différents (Hirst, 1987). La prosodie aurait un rôle central dans la **façon** dont le locuteur produit son énoncé (Cutler et Ladd, 1983; Fonagy, 1991; Levelt, 1989). Nous supposons qu'aux différentes formes illocutoires (FI) correspondraient des formes spécifiques, caractérisées par différents traits acoustiques. Elles feraient partie d'un catalogue de formes prosodiques pré-stockées en mémoire auditive (Sorin, 1989) qui s'apparenterait à un "lexique prosodique", servant de modèle pour leur production et permettant leur reconnaissance. Nous supposons que la construction d'un tel lexique est progressive et que la production de la prosodie est maîtrisée avant sa perception (en accord avec Cutler et Swinney, 1987; Gérard et Clément, 1994).

De plus, dans les actes de parole, les interlocuteurs devraient être capables de prévoir l'évolution des variations des paramètres acoustiques en se référant aux représentations mentales des configurations prosodiques. Plus spécifiquement lors de la production des énoncés, nous supposons que les locuteurs adultes programment la configuration prosodique de l'ensemble de leur énoncé avant l'acte d'énonciation, alors que les enfants gèrent leurs énoncés plus ponctuellement. Nous essaierons de préciser à partir de quel âge les enfants sont capables d'anticiper l'acte productif dès le début de l'énoncé. Sur le plan perceptif, nous supposons que les auditeurs adultes anticipent également l'évolution des variations acoustiques dès le début de la phrase, alors que ces stratégies ne sont pas encore partagées systématiquement par tous les enfants.

Nous avons étudié la production et l'identification de diverses FI grâce à la méthode de *neutralisation sémantique*. Nous présenterons succintement des analyses "globales" de phrase sur le modèle des analyses faites par Gérard et Clément (op. cité), pour délimiter à partir de quel âge les résultats des enfants sont similaires à ceux des adultes. Puis des analyses "locales" ont été effectuées pour vérifier les hypothèses de programmation et d'anticipation. Pour la tâche de production, nous étudierons la durée et la F0 des portions pertinentes dans le signal. Pour la tâche d'identification, nous analyserons les scores de bonnes réponses lors de l'incrémentation mot à mot de la phrase.

2. METHODE

2.1. Production des formes illocutoires

La phrase (P1) "On emmène Michel en vacances" est précédée de mots introducteurs facilitant l'expression des FI : l'affirmation

(I1) est introduite par "maman dit", l'interrogation (I2) par "alors?", la joie (I3) par "chouette", le mécontentement (I4) par "zut". Les 12 locuteurs se répartissent en 4 groupes de 3 enfants de 5, 7 et 9 ans et 3 adultes. La neutralisation sémantique consiste à produire la même phrase selon des FI différentes, afin de se centrer uniquement sur les variations prosodiques. L'ordre de présentation des FI est contrebalancé sur l'ensemble des 12 sujets. Après analyse des énoncés grâce à un éditeur du signal sonore (Unice, Vecsys), les indices globaux suivants portant sur la phrase entière ont été mesurés : 1-durée totale, 2-F0 moyenne, 3-écarts-type et gammes de variation des valeurs de F0, 4-contours mélodiques. Les indices locaux, calculés sur les mots dissyllabiques, sont les suivants : 1-durée de la deuxième syllabe, 2-F0 moyenne par mot.

2.2. Perception des formes illocutoires

Les productions d'un des locuteurs adultes ont été sélectionnées. Après ablation du mot introducteur pour supprimer tout indice sémantique, les énoncés sont présentés aux auditeurs par l'intermédiaire d'un gating particulier, où l'expérimentateur fait écouter successivement mot par mot chacun des énoncés jusqu'à présentation intégrale. Les auditeurs doivent affecter chaque portion de signal par choix forcé à une FI. L'ordre de présentation des énoncés et des FI est contrebalancé sur les 80 auditeurs (20 sujets par âge). Les scores de bonnes détections des FI sont recueillis pour la phrase entière et par portion de signal successif. Les erreurs ont été réparties dans des matrices de confusion.

3. RESULTATS

3.1. Production des formes illocutoires

Le premier résultat intéressant par rapport à notre hypothèse d'un lexique prosodique est que quelque soit l'indice mesuré, globalement ou localement, les FI sont produites systématiquement de façon différenciée.

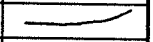
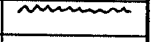
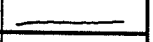
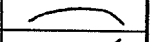
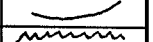
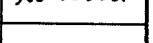
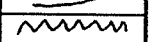
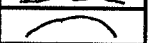
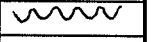
3.1.1. Analyses globales

Nous avons extrait de nos données (cf. table 1) les hiérarchies croissantes des valeurs de durée et de F0 suivantes :

	durée	F0
5 ans	I1<I3<I2<I4	I1<I4<I3<I2
7 ans	I3<I2<I1<I4	I4<I1<I2<I3
9 ans	I2<I1<I3<I4	I4<I1<I3<I2
Ad.	**I2<I1<I3<I4**	**I4<I1<I2<I3**

Les hiérarchies des adultes sont prises comme ordre de référence. Pour la durée, à partir de 9 ans, les productions des enfants s'organisent comme celles des adultes, tandis que pour la F0, cette organisation apparaît dès 7 ans. En ce qui concerne les contours mélodiques, dès 5 ans, les enfants produisent les mêmes types de contours que les adultes, même si les pentes s'accentuent au cours du développement.

Table 1 : Configurations prosodiques des quatre FI observées pour chaque groupe de locuteur : Durées en ms, F0 en Hz, écarts-type des valeurs de F0 (ET), contours mélodiques (CM) et forme des contours mélodiques (récapitulatif des analyses globales de production).

		Durée	F0	ET	CM	Forme
5	*I1*	1801	244,6	41,4	0,52 ↗	
	I2	1994	256,6	43	0,92 ↗	
	I3	1831	252,6	45,2	0,43 ↘	
	I4	2065	245,9	23,5	0,13 ↘	
7	*I1*	1670	242,8	43	0,29 ↘	
	I2	1577	260,3	28,4	0,91 ↗	
	I3	1545	264,1	23,9	0,43 ↘	
	I4	1767	217,7	17,2	0,43 ↘	
9	*I1*	1492	258,5	35,4	0,58 ↗	
	I2	1413	304,3	41,2	0,91 ↗	
	I3	1519	302,3	48,8	0,37 ↘	
	I4	1760	253,5	19,7	0,25 ↘	
A	*I1*	1380	218,6	50,5	0,33 ↘	
	I2	1220	235,2	47,5	0,89 ↗	
	I3	1385	277	60	0,36 ↘	
	I4	1580	190	13,2	0,4 ↘	

3.1.2. Analyses locales

Nous avons analysé chaque mot plein afin de découvrir si l'information expressive se répartit tout au long de l'énoncé. Les hiérarchies suivantes présentent les ordinations des valeurs de durée :

	emmène	Michel	vacances
5 ans	I3<I2<I1<I4	I3<I1<I4<I2	I3<I1<I2<I4
7 ans	I3<I1<I4<I2	I2<I3<I1<I4	I3<I2<I1<I4
9 ans	I2<I3<I1<I4	I2<I3<I1<I4	I1<I2<I3<I4
Ad.	I2<I3<I1<I4	I2=I3<I1<I4	I2<I1<I3<I4

Les adultes et les enfants de 9 ans organisent leur production de la même façon et semblent programmer à l'avance leurs énoncés. Dès le début de la phrase nous retrouvons la même ordination des valeurs par rapport à l'ordre de référence. Par contre les enfants de 5 et 7 ans font des inversions importantes, comme produire l'interrogation avec le débit le plus lent. Les inversions sont moins importantes en fin de phrase à 7 ans.

En ce qui concerne la F0 locale, les hiérarchies sont toutes très semblables dès l'âge de 7 ans, mais les sujets font quelques

inversions systématiques, par rapport à l'ordre de référence.

	emmène	Michel	vacances
5 ans	I1<I4<I2<I3	I4<I1<I2<I3	I1<I4<I3<I2
7 ans	I4<I1<I2<I3	I4<I2<I1<I3	I4<I1<I3<I2
9 ans	I4<I1<I2<I3	I4<I1<I2<I3	I1<I4<I3<I2
Ad.	I4<I1<I2<I3	I4<I2<I1<I3	I4<I1<I3<I2

Pour l'indice temporel, l'allongement peut se distribuer tout au long de l'énoncé; tandis que pour l'indice intonatif, la variation des valeurs de F0 se fait en fonction de la forme des contours mélodiques, ce qui expliquerait la plupart des modifications d'ordre. Nous pouvons donc penser que ces sujets programment la tension de leurs cordes vocales pour placer adéquatement le ton tout au long de l'énoncé.

3.2. Perception des formes illocutoires

3.2.1. Analyses globales

Les scores de bonnes détections sont minoritaires chez les enfants de 5 ans (33.5%) mais supérieurs à une répartition au hasard (20%), augmentent à 7 ans (48%) et sont proches pour les enfants de 9 ans (71%) et les adultes (77.5%).

3.2.2. Analyses locales

Si l'information est insuffisante, si les anticipations de l'évolution future du signal sont impossibles et si la décision est vraiment aléatoire, les pourcentages de bonnes réponses devraient osciller autour de 20%. Nous avons considéré que 40% (cf. table 2) pouvait être signe d'un traitement cognitif d'anticipation, à condition que ce score ne soit pas isolé. D'après ce double critère d'analyse, les enfants de 5 ans ne traitent jamais les énoncés de façon anticipée. Par contre, certains enfants de 7 et 9 ans commencent à identifier correctement les FI avant la fin de la phrase. Le mécontentement est identifié à partir de l'écoute de trois mots et l'affirmation à partir de quatre mots à 9 ans. Au total, 3 pourcentages non isolés sont supérieurs ou égaux à 40% pour les enfants de 7 ans. Les anticipations sont à peine plus fréquentes à 9 ans avec 5 valeurs, par contre dès l'écoute du dernier mot, ces enfants ne se trompent quasiment plus. Les scores des adultes montrent des anticipations nombreuses de l'affirmation, de l'interrogation et du mécontentement. Nous considérons que seuls les adultes (13 valeurs) font réellement des anticipations correctes de l'évolution future des signaux sonores, à tout le moins pour P1.

3.2.3. Analyse des choix erronés

Nous avons étudié les matrices de confusion suivantes : répartition des erreurs lors de l'écoute des phrases complètes, sans le dernier mot, sans les deux derniers,... Le choix de la forme affirmative est l'erreur la plus fréquente à tout âge lors de l'écoute de la phrase complète; mais si nous enlevons le dernier mot - porteur du plus grand nombre de contrastes prosodiques - le choix erroné de l'interrogation devient important. Les enfants de 5 ans montrent une indifférenciation dans les choix erronés, tandis que les autres sujets semblent faire des erreurs systématiques. Ils confondent plus souvent deux FI si celles-ci se ressemblent acoustiquement (cf. les études de production). Les confusions les plus importantes se font par exemple entre le mécontentement (présenté) et l'affirmation (choisie), entre la joie et l'interrogation.

Table 2 : Pourcentages d'identification correctes des FI par portions successives du signal : On (M1) + emmène (M2) + Michel (M3) + en (M4) + vacances (M5), (analyses locales de perception).

5	**I1**	**I2**	**I3**	**I4**
M1	10	10	30	20
M2	35	20	25	25
M3	35	10	35	50
M4	25	10	10	25
M5	**50**	**20**	**35**	**65**
7	**I1**	**I2**	**I3**	**I4**
M1	25	15	10	20
M2	35	40	20	15
M3	30	30	55	55
M4	30	20	25	50
M5	**65**	**55**	**25**	**95**
9	**I1**	**I2**	**I3**	**I4**
M1	40	25	35	30
M2	35	40	45	35
M3	30	40	35	60
M4	40	15	30	50
M5	**80**	**85**	**35**	**100**
A	**I1**	**I2**	**I3**	**I4**
M1	65	25	35	10
M2	60	45	40	80
M3	45	50	20	90
M4	40	45	10	90
M5	**100**	**100**	**55**	**100**

DISCUSSION

La durée et la F0 semblent s'organiser en configurations prosodiques d'ensemble, mais cette dépendance entre les paramètres acoustiques correspond au stade achevé du développement. Il semblerait qu'un allongement de la durée s'accompagne en général d'une baisse de la hauteur de la voix, qu'un raccourcissement s'accompagne d'une élévation dans les tons aigus et qu'une durée

moyenne soit produite avec une hauteur moyenne.

L'étude développementale a montré que la maîtrise des paramètres prosodiques dans l'expression des FI ne s'acquiert pas en une seule étape. La production de contours mélodiques différenciés semble être acquise assez tôt dès l'âge de 5 ans. A 7 ans, les enfants commencent à placer la hauteur de leur voix comme les adultes en fonction des FI, et à 9 ans s'y ajoute le contrôle du débit d'élocution. Les adultes produisent toujours différemment les FI sur la base de tous les indices analysés, et surtout programment dès le début de l'énoncé la hauteur (en tenant compte de la pente des contours mélodiques), et modifient la durée en distribuant les allongements ou raccourcissements tout au long de l'énoncé.

Les formes expressives, comme la joie, ne sont pas produites par les enfants de 5 ans dans cette situation de simulation. On sait que les enfants expriment spontanément leurs émotions, mais, en référence aux théories de l'esprit de Flavell, il n'est pas exclu que les enfants de cet âge soient incapables de *simuler* ce genre d'émotions et nos résultats pourraient provenir essentiellement de la procédure choisie.

Enfin, à partir de l'ensemble de nos résultats, il nous semble que l'existence d'un lexique prosodique soit justifiée pour plusieurs raisons. Lors de la production des FI, les locuteurs semblent bien faire référence à des représentations mentales stockées en mémoire, puisqu'ils produisent les mêmes configurations prosodiques dès 7 ans pour l'intonation, 9 ans pour la maîtrise complète des paramètres prosodiques. De même, sur le plan perceptif, les auditeurs se trompent peu à partir de 9 ans et identifient correctement les FI. Lorsque les sujets sont dans une situation de doute, il semble qu'ils recherchent en mémoire les représentations adéquates, et leurs erreurs proviennent soit du choix d'une forme de référence (affirmation ou interrogation avant le déroulement complet de l'énoncé), soit du choix d'une forme ressemblant acoustiquement à l'énoncé présenté. De telles données évoquent un concept utilisé pour le lexique mental, l'idée de gêne entre les voisins prosodiques. Gérard et Clément (1994) montraient que les adultes et les enfants de 7 ans partageaient les mêmes invariants structuraux globalement pour la production et à partir de 10 ans pour la perception. Ceci est à nuancer grâce à cette étude développementale : les différents paramètres ne sont pas maîtrisés en même temps, la construction du lexique prosodique, s'il existe, est donc progressive, et ses invariants structuraux ne seraient maîtrisés "complètement" qu'à partir de 9 ans. Ceci est également à nuancer par les phénomènes de programmation et d'anticipation : lors de la tâche de production, les enfants programment à l'avance l'énonciation des FI dès qu'ils savent gérer un paramètre prosodique, alors qu'a priori, nous pensions que seuls les adultes en étaient capables; tandis qu'au cours de la tâche de perception, même à 9 ans, les enfants ont du mal à anticiper l'évolution des variations des paramètres prosodiques avant la fin du déroulement complet des énoncés. Si les mêmes représentations mentales président à la production et à la perception des FI, les chemins d'accès doivent être différents, car les types de traitements cognitifs sous-jacents à l'énonciation et à l'identification sont décalés au cours du développement. Ainsi, le lexique prosodique semble acquis dans sa totalité à l'âge de 9 ans pour les deux types de traitements requis à la communication, mais toutes les finesses du traitement en temps réel des énoncés ne sont pas totalement maîtrisées lors de la perception des FI.

5. BIBLIOGRAPHIE

Caron, J. (1983). *Les régulations du discours: psycholinguistique et pragmatique du langage.* Paris, Presses Universitaires de France.

Caron, J. (1989). *Précis de psycholinguistique.* Paris, Presses Universitaires de France.

Cutler, A. & Ladd, D.R. (1983). *Prosody : Models and measurements.* Springer-Verlag, Berlin, Heidelberg. p.1-9.

Cutler, A. and Swinney, D.A. (1987). Prosody and the development of comprehension. *Journal of child language*, **14**, p.145-167.

Fonagy, I. (1991). *La vive voix, essais de psycho-phonétique.* Bibliothèque scientifique Payot.

Gérard, C. et Clément, J. (1994). Patrons prosodiques et intentions des locuteurs: production et perception de formes expressives chez l'adulte et l'enfant, *Actes des 20ès Journées d'étude sur la parole*, Lannion, France, p.57-62.

Hirst, D. (1987). Intonation, syntaxe, sémantique et pragmatique. *SIGMA (Publication du C.E.L.A.)*, **11**, p.148-170.

Levelt, W.J.M. (1989). *Speaking. From intention to articulation.* Cambridge, Massachusetts: The MIT Press.

Sorin, C. (1989). Perception de la parole continue, in M.C. Botte, G. Canévet, L. Demany et C. Sorin, *Psychoacoustique et perception auditive*, série audition, INSERM/SFA/CNET, p.123-139.

SYNCHRONISATION DU NIVEAU TONAL SUR LE NIVEAU SEGMENTAL EN LECTURE : ETUDE PRELIMINAIRE

Pascale NICOLAS, Daniel HIRST

Institut de phonétique d'Aix-en-Provence- Laboratoire "Parole et Langage" URA 261, CNRS
29, Av.R.Schuman, 13621 Aix-en-Provence
Tél. : 42 95 36 34 - Fax : 42 59 50 96 - e-mail : {nicolas,hirst}@lpl.univ-aix.fr

ABSTRACT

In a previous study on the intonative organisation of read text in French, we developped a frequential coding of target points given by a modelling algorithm of the fundamental frequency. In the present study, we wish to enlarge our analysis to the temporal behavior of the coded target points with respect to the vowel onsets.

1.INTRODUCTION

Dans deux études précédentes (Nicolas et Hirst, 1995; Nicolas, 1995) nous avons analysé les phénomènes intonatifs propres à la lecture, afin de rechercher la possible influence sur l'organisation intonative d'une structure telle que le texte lorsqu'il est constitué de plusieurs paragraphes. Ces analyses ont été réalisées à partir d'une modélisation par points cibles et de leur codage par un système symbolique INTSINT (INternational Transcription System for INTonation, Hirst et Di Cristo, à paraître). Elles nous ont permis d'effectuer, dans un premier temps, l'analyse des évolutions des courbes intonatives sur un axe fréquentiel. Mais nous pensons que l'analyse du comportement temporel des points cibles codés par rapport à la chaîne segmentale représente également une étape obligatoire afin d'établir une codification de leur emplacement temporel.

Les problèmes de synchronisation des événements intonatifs sur le segmental ont été le plus souvent abordés dans le domaine de la synthèse de la parole (Rietveld et Gussenhoven, 1995; Kohler, 1991). Ainsi une synchronisation inadéquate, même si les formes intonatives sont correctes, peut conduire à de mauvais résultats (effet non "naturel" ou encore changements de sens inattendus) quant à la parole de synthèse obtenue.

Des études (Rietveld et Gussenhoven, 1995, van Santen et Hirschberg, 1994) ont montré que la composition segmentale des syllabes accentuées (le nombre et la durée des éléments ainsi que la présence de consonnes sonores) influençait la localisation des pics de fréquence fondamentale. Dans cette étude, nous ne partons pas du tout d'une typologie accentuelle, mais simplement d'une typologie tonale qui définit les variations mélodiques. Nous choisissons identiquement aux études (Nicolas et Hirst, 1995; Nicolas, 1995) réalisées sur les comportements fréquentiels des points cibles, d'opter pour une observation globale des phénomènes de synchronisation tonale/segmentale tout au long d'une lecture de texte, sans se référer ni à la structuration syllabique, ni à leur caractère accentué. Ainsi, nous choisissons les attaques vocaliques comme lieu de "rendez-vous" entre le niveau intonatif et le niveau segmental. Nous pensons que ce lieu constitue un point d'ancrage satisfaisant pour une étude exploratoire. Nous cherchons avant tout à savoir si la méthodologie employée nous permet de capter des phénomènes intéressants de synchronisation.

2.MATERIEL ET METHODOLOGIE

Le corpus utilisé dans cette étude consiste en un texte composé de trois paragraphes totalisant 8 phrases. Le corpus mis à disposition des quatre locuteurs natifs de langue française (2 femmes et 2 hommes) conservait les marques typographiques du texte d'origine. Ce texte possède une structure discursive unique et stable, ce qui revient à inhiber l'influence de changements de sens sur la production de l'intonation. Les enregistrements analogiques ont été réalisés en chambre anéchoïque, puis numérisés à 16 kHz (résolution 16 bits). On effectue ensuite la détection de fréquence fondamentale ainsi que sa modélisation automatique (Espesser, 1993; Hirst et Espesser, 1993). Cette modélisation est représentée par une séquence de points cibles sous la forme de couples de valeurs <F0, temps>. Les points cibles correspondent aux variations locales pertinentes de la courbe mélodique. Ils permettent, reliés par une fonction d'interpolation (du type "quadratic spline"), de retrouver le profil suprasegmental (où n'interviennent pas les variations

micromélodiques) caractérisant globalement l'intonation du texte lu. Les erreurs de modélisation ont été corrigées manuellement avec une validation auditive. Une macro-segmentation du continuum intonatif est réalisée essentiellement sur la base de la ponctuation présente dans le texte écrit. Cependant, nous prenons également en compte les connecteurs et les pauses de plus de 550ms réalisées sans la présence de ponctuation. Pour toutes les unités ainsi définies, chaque point cible va recevoir un symbole du système de transcription de l'intonation INTSINT. On détermine les points de niveau absolu à l'intérieur de chaque unité : le premier point cible recevra le symbole M(id) (peut être dans certain cas T ou B), le point cible le plus haut T(op), le point cible le plus bas B(ottom). Les autres points cibles appartenant à l'unité auront une codification à niveau relatif et suivront le principe de la comparaison successive. Nous coderons chaque point cible comme H(igh) (pic), L(ow) (creux), D(ownstep) (baisse graduelle) ou U(pstep) (hausse graduelle) en fonction de la valeur fréquentielle du point cible suivant et précédent. Cette codification nous donne l'opportunité d'analyser en fonction de ces symboles le comportement temporel des points cibles tout au long d'une lecture de texte.

Pour le calcul des distances, nous ne prenons pas en compte les schwas comme élément moteur d'une structure définie par les attaques vocaliques. Nous calculons les distances séparant les points cibles des attaques vocaliques (figure 1) à l'intérieur de chaque unité telle que nous l'avons définie précédemment. De ce fait, certaines distances n'ont pas été prises en compte lorsque aucune attaque vocalique ne suivait ou ne précédait le point cible à l'intérieur de l'unité. Aucune pause n'est incluse dans le calcul des distances. A l'examen des distances ainsi calculées, celles du locuteur C semblent avoir un comportement très différent. Ce locuteur est un lecteur professionnel qui utilise abondamment les pauses et les changements de tempo pour structurer sa lecture. Il possède globalement un débit de parole assez lent (les distances sont en moyenne importantes). Nous relevons également une distance précédant le point cible de l'attaque vocalique plus importante que la distance qui suit (seul cas parmi nos locuteurs). Le débit semble donc influencer la synchronisation des points cibles avec les attaques vocaliques. Si la distance suivant le point cible jusqu'à l'attaque vocalique ne semble pas trop touchée par le changement de débit (comportement identique pour les quatre locuteurs), la distance précédente semble par contre subir des variations.

Nous prenons alors la décision de ne conserver que 3 de nos locuteurs (possédant un débit moyen comparable) en gardant en mémoire que des tests complémentaires sur la synchronisation doivent être réalisés afin d'analyser les effets des différents débits (très rapide ou très lent) sur la synchronisation des points cibles avec la chaîne segmentale. De plus, puisque les trois locuteurs restants ont le même comportement, nous évitons ainsi d'inclure cette variable indépendante dans les analyses statistiques.

3. RESULTATS

Nous explorons isolément les deux types de distances précédant et suivant les points cibles jusqu'aux attaques vocaliques (457 distances précédentes et 428 distances suivantes calculées pour les trois locuteurs, table 1). Nous prenons en compte également un autre type de calcul nous permettant d'analyser le comportement temporel des points cibles entre deux attaques vocaliques. Les mesures que nous avons effectuées ont évidemment des écarts types assez importants.

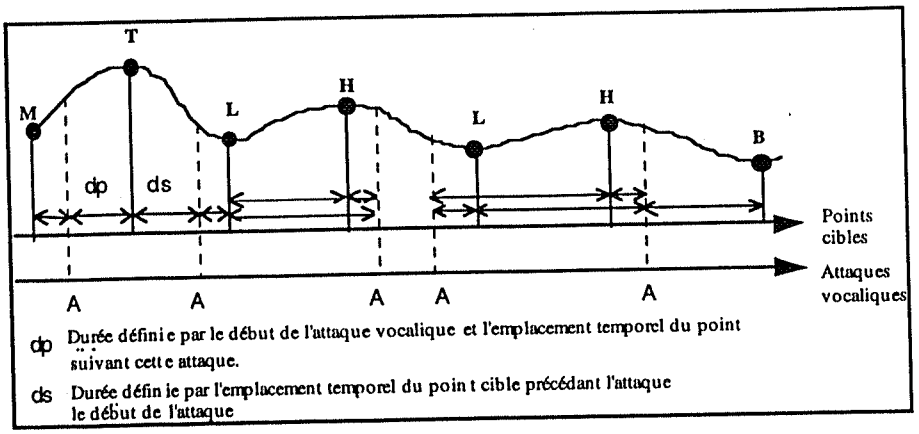


Figure 1 : Calcul des distances séparant le point de l'attaque vocalique précédente (dp) et des distances séparant le point cible de l'attaque vocalique suivante (ds) pour une unité stylisée.

3.1. Distance précédente :

Cette distance s'avère significative en fonction des symboles attribués aux points cibles (p=,0019). Le premier point cible de l'unité qui est essentiellement défini par un critère de localisation (premier point de l'unité) possède une distance précédente faible (moy=46ms) qui se détache de manière significative des distances obtenues pour les autres symboles. Ce point cible n'est pas censé représenter quelque chose de linguistiquement déterminant au niveau de son chronométrage par rapport au segmental. Par contre, il est certain que sa composante fréquentielle est très importante. Cette hauteur dépend généralement de la hauteur fréquentielle du dernier point cible de l'unité précédente, mais aussi de la longueur de la pause qui précède. Il n'est donc pas très intéressant dans une étude portant sur les phénomènes de synchronisation temporellle d'analyser le comportement fréquentiel de ce type de symbole. Le point cible B(ottom) se différencie également par la distance (moy=132ms, distance la plus longue) qui le précède par rapport à l'attaque vocalique. Lorsque l'on regarde le comportement de cette distance par rapport aux distances des autres points cibles, elle se différencie des symboles T, H et D. Il semble donc que le point cible le plus bas de l'unité soit retardé par rapport aux points représentant des pics de fréquence fondamentale et par rapport à une baisse graduelle de fréquence fondamentale (symbole D(ownstep)).

3.2. Distance suivante :

Lorsque l'on analyse la distance séparant chaque point cible de l'attaque vocalique suivante, on obtient un facteur p=0,0001.

Les points cibles T (moy=164ms) et H (moy=136ms) possèdent en moyenne les plus grandes distances. Le point cible portant le symbole T se différencie de tous les autres symboles. Tandis que les points cibles codés avec le symbole H se détachent de tous sauf de U. Les points cibles correspondant à des pics de fréquence fondamentale (codés T ou H) semblent donc posséder une localisation temporelle plus en avance que les autres symboles en regard de la localisation de l'attaque vocalique suivante.

3.3. Localisation du point cible entre deux attaques vocaliques :

Grâce à ce calcul, nous pourrons examiner le comportement de la localisation du point cible par rapport à celles des attaques vocaliques adjacentes. Si la valeur est inférieure à 0.5, alors le point cible sera plus proche de l'attaque vocalique précédente, si la valeur est supérieure à 0.5 alors le point cible sera plus proche de l'attaque vocalique suivante.

Nous obtenons un facteur de probabilité p=0,0002 pour que ces valeurs soient significativement différentes en fonction du symbole du point cible. Puisque la distance suivante est en moyenne plus importante que la distance précédente (comportement suivi par tous les locuteurs), on a le plus souvent des valeurs inférieures à 0,5 (figure 2). Bien que les effets simples ne s'avèrent pas tous significatifs, nous trouvons que le symbole T(op) (moy=0,327) se différencie des symboles représentant des creux ou des configurations descendantes (B, L et D) de fréquence fondamentale. Ainsi par rapport aux attaques vocaliques, T est le symbole le plus à gauche (à

Table 1 : Moyenne (en ms), écart type et effectif pour chaque symbole et pour l'ensemble des locuteurs en fonction des différentes mesures : distance du point cible à l'attaque précédente, distance du point cible à l'attaque suivante et pour le calcul correspondant à la distance précédente divisée par la distance suivante ajoutée à la distance précédente.

Distances	Distance précédente			Distance suivante			ds/(ds+dp)		
Points cibles	Moy	ET	N	Moy	ET	N	Moy	ET	N
M	46,2	54,5	26	101,7	75,6	46	0,258	0,256	24
T	100,9	73,1	56	164,1	74,9	45	0,327	0,205	45
B	132,7	90,7	55	101,6	75,6	32	0,5	0,25	31
H	102,6	78	129	136,1	94,7	111	0,409	0,303	108
L	113,6	94,1	108	99,6	58,5	111	0,488	0,291	107
U	106,5	98,7	18	113,8	90,2	19	0,461	0,332	18
D	97,4	76,1	65	99,3	99,3	64	0,513	0,304	63

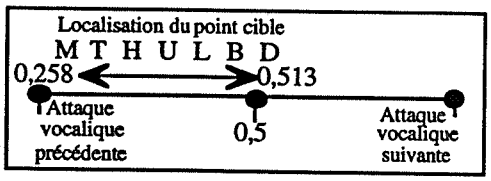


Figure 2 : Localisation des points cibles par rapport aux attaques vocaliques précédentes et suivantes.

l'exception du symbole M). Le symbole H vient ensuite (moy=0,409), mais cet effet ne le différencie pas de B. Par contre, il est intéressant de remarquer que H se différencie significativement de L et de D.

4. DISCUSSION

Par rapport à la distance précédente, seuls les symboles M et B semblent posséder un comportement particulier. Le premier possède une distance précédant le point cible de l'attaque vocalique très faible alors que le second a une distance importante. Ils correspondent la plupart du temps, pour le codage des points cibles à l'intérieur des unités, à des tons de frontières. Ce qui expliquerait leur localisation précoce ou tardive. Il apparaît également que les points cibles sont en majorité (figure 2) plus proches de l'attaque vocalique précédente que de l'attaque vocalique suivante. Nous avons soulevé en début de ce travail, que pour le locuteur qui possède un débit lent (analysé à part dans cette étude), nous avions une redistribution des distances précédentes en fonction des points cibles codés. Notamment, le symbole T possédait une distance très importante (à l'inverse des autres locuteurs), nous montrant que ce type de pic (dans une moindre mesure pour le pic H) pouvait être retardé par rapport à l'attaque vocalique précédente pour se trouver dans une position beaucoup plus centrale par rapport à l'unité définie par les deux attaques. Un changement de débit ne semble donc pas avoir une influence uniforme sur les deux types de distance. Ceci peut tout à fait remettre en question ce que nous avons observé pour nos trois locuteurs, c'est à dire que les pics sont globalement plus proches par rapport à l'attaque vocalique qui précède que les creux de fréquence fondamentale.

En conclusion de cette étude préliminaire, nous pouvons dire qu'avec la méthodologie choisie (abstraction des phénomènes accentuels et de la structuration syllabique), nous pouvons retrouver des phénomènes constants de synchronisation des phénomènes fréquentiels et segmentaux. Cependant, certains de ces phénomènes sont modifiés par les changements de débit. Dans le cas de la lecture de texte, nous relevons de nombreux phénomènes de changements de débit (première phrase du paragraphe rapide, ralentissement en fin de paragraphe et fin de texte). De prochaines études devront être réalisées sur le comportement des points cibles codés H et T (Silverman et Pierrehumbert, 1990) en fonction de la variation de débit des unités que nous avons définies tout au long du texte. Nous pensons que cette étape s'avère nécessaire avant d'aller plus en avant dans notre recherche, c'est à dire l'exploration de ce qui se produit au niveau de la syllabe et au niveau de l'accentuation.

5. BIBLIOGRAPHIE

Nicolas, P., Hirst, D. J. (1995), "Symbolic coding of higher-level characteristics of fundamental frequency curves", *Proceedings of the 4th Eurospeech*, Madrid, Spain, Volume 2, 989-992.

Nicolas, P. (1995) *Organisation intonative du texte lu en français*, Thèse nouveau régime, Aix-en-Provence, 349.

Hirst, D.J.; Di Cristo, A. (à paraître), "A survey of intonation systems" in *Intonation systems : a survey of twenty languages,* Cambridge University Press.

Rietveld, T.; Gussenhoven C. (1995) "Aligning pitch targets in speech synthesis : effects of syllable structure", *Journal of Phonetics,* n°23, 375-385.

Kohler, K.J. (1991), "Prosody in speech synthesis : the interplay between basic research and TTS application", *Journal of Phonetics,* n°19, 121-138.

van Santen J., Hirschberg J. (1994), "Segmental effects on timing and height of pitch contours", ICSLP 94, Yokohama, S14-3.1, 719-722.

Espesser, R. (1993) "MES : Motif Editeur de Signal", document interne, laboratoire "Parole et Langage" URA 261, CNRS.

Hirst, D.; Espesser, R. (1993) "Automatic modelling of fundamental frequency using a quadratic spline function", *Travaux de l'Institut de Phonétique d'Aix* n°15, .71-85.

Silverman, K.; Pierrehumbert, J. (1990), "The timing of prenuclear accent in English", in *Papers in Laboratory Phonology*, 72-106.

UN MODELE CONNEXIONISTE MODULAIRE POUR L'APPRENTISSAGE DES « GESTES » INTONATIFS

Yann Morlec, Gérard Bailly et Véronique Aubergé

Institut de la Communication Parlée, INPG & Université Stendhal
46, av. Félix Viallet 38031 Grenoble Cedex 01, France
e-mail : (morlec, bailly, auberge)@icp.grenet.fr

1. ABSTRACT

A dynamic model for synthesizing intonation is presented. This model is based on the following assumptions: intonation is the result of superposed and independent prototypical gestures belonging to the diverse linguistic levels: sentence, proposition, syntagm, group, word and phoneme. Prototypical movements are progressively stored in a prosodic lexicon and used by the speaker in given communication tasks. Our current implementation of this model is a modular association of sequential neural networks (SNNs). Each SNN is in charge of the melodic prediction of a specific linguistic level. The resulting melody is the weighted sum of SNNs outputs. We presently focus on the sentence level. We built a corpus of 1932 sentences pronounced with 6 attitudes and various lengths. We then designed SNNs able to perform the expansion of prosodic sentence movement. Preliminary results show that these simple SNNs can give acceptable F0 prediction and keep essential features of each attitude whatever the syllabic length of the sentence.

2. INTRODUCTION

Nous présentons ici une stratégie d'apprentissage d'un modèle dynamique de génération de l'intonation du Français. Notre approche théorique postule que la courbe mélodique est la superposition de « mouvements » intonatifs correspondant aux différents niveaux de représentation du discours (la phrase, la proposition, le syntagme, le groupe, le mot et le phonème). La synthèse de l'intonation résulte de la somme pondérée des prédictions de réseaux connexionistes récurrents modélisant l'expansion des « gestes » prosodiques associés à chacun de ces niveaux. Une attention particulière est portée à la modélisation des contours de phrase véhiculant notamment la position du locuteur vis-à-vis de son discours.

3. CADRE THEORIQUE

Notre analyse de la mélodie est guidée par une approche morphologique. Nous pensons que l'auditeur a des attentes sur ce qu'est une intonation « normale », comme il existe une syntaxe « normale » et qu'il peut ainsi interpréter comme une mise en relief quelque chose qui émerge du fond, qui dévie d'une forme attendue. Cette notion d'attente est à rapprocher aux propositions faites par le mathématicien Petitot-Concordat (Petitot-Concordat, 1986) élève de R. Thom : l'analyse morphologique est un lieu privilégié où notre perception du monde physique peut être guidée - ou biaisée -

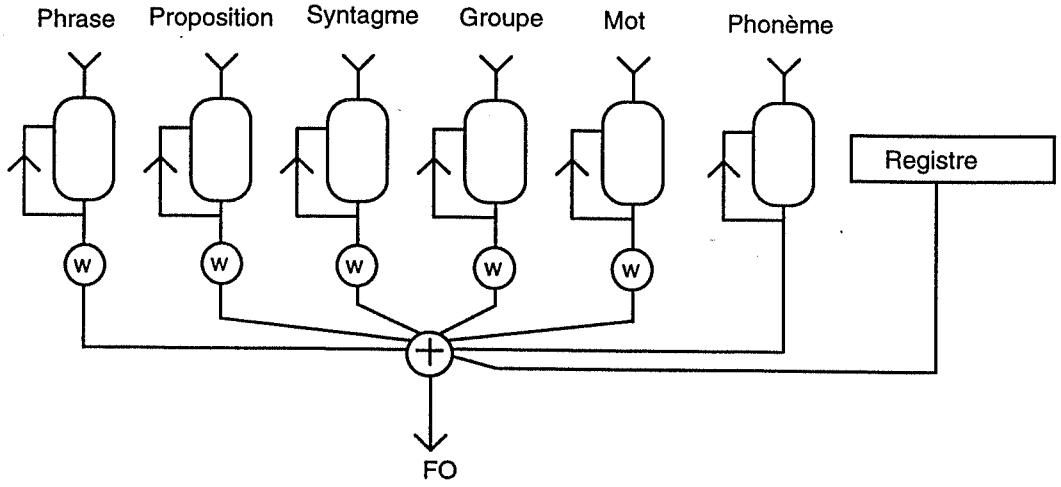


Figure 1: Architecture du modèle de génération des « gestes » intonatifs. Les sorties de chaque module sont exprimées en 1/4 de tons. Le registre donne la fréquence laryngienne moyenne. Les pondérations w permettent d'ajuster l'influence des contours portés.

par la projection de nos représentations mentales préexistantes. C'est par cette boucle action-perception-cognition que nos représentations mentales émergent et s'enrichissent.

Dans cette optique, nous considérons que chaque niveau de description linguistique sur l'axe syntagmatique génère des indices voire des formes qui lui sont propres et qui maintiennent entre elles des contrastes propres à ce niveau. Ainsi la ligne mélodique observée est le résultat de la superposition de ces formes. Nous considérons ici les niveaux de représentation suivants : la *phrase* dont le contour mélodique va véhiculer l'attitude du locuteur. Ce contour de phrase, « geste » porteur, est modulé successivement par les contours de *proposition*, de *syntagme* et de *groupe* qui permettent, en particulier, de rendre compte du degré de cohésion entre les différents constituants de ces niveaux. La micromélodie est l'étape ultime de cette modulation hiérarchique et additive.

L'étude de modèles de superposition pour la génération de l'intonation n'est pas nouvelle. Thorsen propose un modèle ou la ligne mélodique est obtenue par addition de segments de droite (Thorsen, 1980). Fujisaki la décrit comme la superposition de réponses de filtres (Fujisaki, 1971). Notre approche diffère de ces deux exemples par le fait que nous n'imposons pas a priori la forme des contours intonatifs. Par exemple, certains « mouvements » mélodiques de phrase peuvent être complexes et sujets à de brusques variations. De cette description de l'intonation découle le modèle de génération décrit ci-après.

4. UN MODELE D'EXPANSION DU « MOUVEMENT » MELODIQUE

Le modèle dynamique de génération de l'intonation, dans sa version actuelle, consiste en une association modulaire de réseaux connexionistes récurrents chargés de prédire les « gestes » intonatifs des niveaux linguistiques qui leur sont associés. Le contour mélodique final est obtenu par la somme pondérée de ces prédictions (figure 1).

4.1 Des « mouvements » superposés

L'apprentissage du modèle de génération se fait de manière itérative (Aubergé, 1992) - du niveau porteur le plus large vers le niveau porté le plus petit - par la gestion dynamique des pondérations. A la manière des stratégies observées d'apprentissage supervisé de l'intonation (Strömqvist & al., 1995), le modèle va d'abord apprendre à générer un contour mélodique global de phrase. Pour ceci, tous les poids *w* sont initialisés à zéro sauf celui de phrase. Lorsqu'il a acquis une performance suffisante, il va alors moduler ce « geste » mélodique de phrase par l'ajout de la contribution des niveaux inférieurs ; ceci en gelant les paramètres des modules déjà appris et en positionnant le poids *w* correspondant à 1.

La modularité de cette architecture permet donc d'utiliser un corpus approprié à chaque niveau linguistique étudié : le modèle d'expansion du « mouvement » de phrase pour diverses attitudes peut être appris sur un corpus de phrases spécifiques constitué par exemple de mono-mots (cf. paragraphe suivant) alors que, parallèlement, l'apprentissage de la typologie des « mouvements » de groupes peut être effectué sur des phrases assertives à fort branchement syntaxique (Aubergé, 1992).

4.2 Des « mouvements » pondérés

L'existence de pondérations sur les différents niveaux se justifie par les observations suivantes :

- Pour les attitudes correspondant à une reprise mot pour mot de l'assertion de l'interlocuteur (*doute-incrédulité, ironie de soupçon),* le locuteur ne ressent pas le besoin de segmenter à nouveau cet énoncé. La réalisation mélodique se restreint alors à un contour de phrase (pondération nulle sur le niveau groupe). Par contre, dans le cas d'une exclamative, la segmentation des contours portés est largement exagérée (pondération supérieure à 1).

- La gestion des poids permet de contrôler dynamiquement l'importance relative des niveaux. Ainsi notre modèle gère facilement les variations de pente de la ligne de déclinaison observées dans des tâches de lecture de textes (Silverman, 1987).

Nos recherches se portent actuellement sur le niveau de la phrase, le découpage de l'énoncé ayant déjà été largement étudié (Morlec & al., 1995).

5. LE NIVEAU PHRASE

Nous nous intéressons ici à la manière par laquelle les locuteurs véhiculent par la voix diverses attitudes. Le but étant de démontrer qu'ils produisent des formes mélodiques bien identifiables.

5.1 Elaboration du corpus

Le corpus *attitudes* comporte 322 phrases de longueur variable (1 à 8 syllabes) prononcées suivant six attitudes intonatives différentes référencées dans (Calbris & al., 1981), soit un total de 1932 phrases. La longueur réduite

des phrases du corpus permet de limiter l'influence des contours portés sur la courbe mélodique.

5.2 Caractérisation phonétique

Les réalisations mélodiques sont caractérisées par trois valeurs de F0 par syllabe. Ces valeurs sont prélevées sur la voyelle : on considère qu'en Français, la syllabe ne reçoit qu'un seul mouvement stylisée par une fonction polynomiale du second ordre.

5.3 Description des attitudes

Parmi les six attitudes retenues, deux s'apparentent à une *intonation d'énoncé* où les sentiments personnels du locuteur à l'égard du message qu'il émet sont occultés : l'*affirmation* et la *question simple*. Les quatre autres sont des *intonations de discours* faisant apparaître des informations sur la situation dans laquelle se trouvent les interlocuteurs, sur les dispositions du locuteur vis-à-vis de l'objet dont il parle : l'*exclamation de surprise*, le *doute-incrédulité*, l'*ironie de soupçon* et l'*évidence*. Le *doute-incrédulité* exprime le désaccord partiel et à la limite une négation de ce qui a été préalablement affirmé. Sa mélodie est surtout caractérisée par une mise en relief de l'avant dernière syllabe suivie d'une dernière syllabe montante. Dans l'*ironie de soupçon*, on reprend l'affirmation de l'interlocuteur en lui faisant comprendre qu'on la met en doute, qu'on la contredit. La prosodie de ce type de phrase est caractérisée par un ton haut et, brusquement, sur la dernière syllabe, par une chute de fondamental.

La figure 2 montre la superposition de contours mélodiques (mono-mots essentiellement) pour des phrases de une à huit syllabes correspondant à l'attitude de *doute-incrédulité*. Ces courbes, comme celles associées aux autres attitudes étudiées, donnent les caractéristiques essentielles de l'expansion du « mouvement » prosodique en fonction du nombre de syllabes. Ces contours paraissent se diviser en deux parties : une phase finale de capture très prototypique sur une à deux syllabes précédée d'une phase d'amorce caractérisée par une valeur initiale assez stable et suivie d'une relaxation.

Cette analyse met en lumière l'existence de contours mélodiques globaux qui ne sont pas « ancrés » sur les niveaux inférieurs de la phrase. Pour les intonations d'énoncé, on peut néanmoins remarquer l'apparition de modulations supplémentaires lorsque la phrase dépasse typiquement 5 à 6 syllabes. Cette modulation est due, dans notre modèle, à la nécessité d'apporter par l'adjonction de contours portés des informations sur découpage de l'énoncé en groupes.

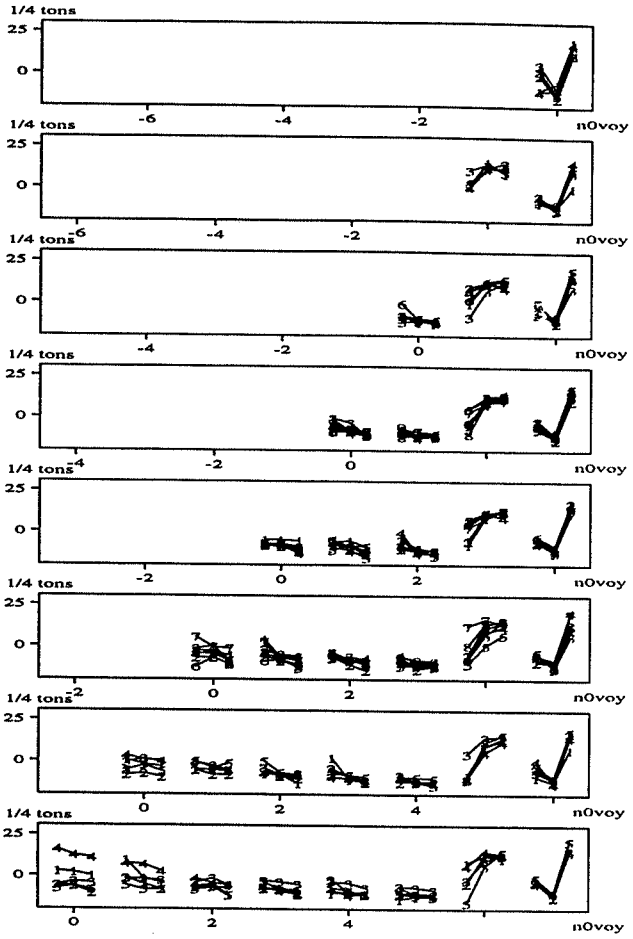


Figure 2 : Contours mélodiques pour l'attitude de doute-incrédulité appliquée à des phrases de 1 à 8 syllabes.

5.4 Architecture du module *phrase*

Il comporte un ensemble de réseaux de neurones de type Elman (Elman, 1988) destinés à prédire les contours mélodiques véhiculant les six attitudes enregistrées dans le corpus *attitudes*.

Ces réseaux dynamiques ont une architecture semblable comportant :

- En entrée, deux rampes linéaires décroissantes relative (de 1 à 0) et absolue (du nombre de syllabes à 0). Elles permettent de gérer la séquentialité du « mouvement » et de déclencher la phase de capture au bon moment.
- Deux couches cachées interconnectées dans les deux sens.
- La sortie correspond à la F0 stylisée.

5.5 Apprentissage

Les corpus d'apprentissage de ces réseaux sont constitués de l'ensemble des phrases mono-mot de 1 à 6 syllabes, soit 18 phrases par attitude (les mono-mots permettent de neutraliser l'incidence des contours portés). Les phrases de 7 et 8 syllabes sont, quant à elles, destinées à vérifier leur capacité de généralisation (corpus de test).

6. RESULTATS

Sur le corpus d'apprentissage, les prédictions générées par le module *phrase* sont proches des moyennes des échantillons d'apprentissage (comprises entre $\mu-\sigma$ et $\mu+\sigma$).

Pour l'attitude de *doute-incrédulité*, les figures ci-après permettent de comparer les prototypes prédits par le réseau (figure 4) avec la moyenne des contours d'apprentissage (figure 3). Notez que le réseau a su reproduire fidèlement les caractéristiques essentielles de ce type d'attitude : on retrouve la mise en relief de l'avant-dernière syllabe et le « mouvement » montant sur la dernière.

La généralisation des réseaux à été étudiée sur les phrases du corpus de test de 7 et 8 syllabes (figure 4). Les résultats montrent que la phase de capture est toujours bien mémorisée. L'expansion du « mouvement » se réalise correctement sur l'étape de préparation.

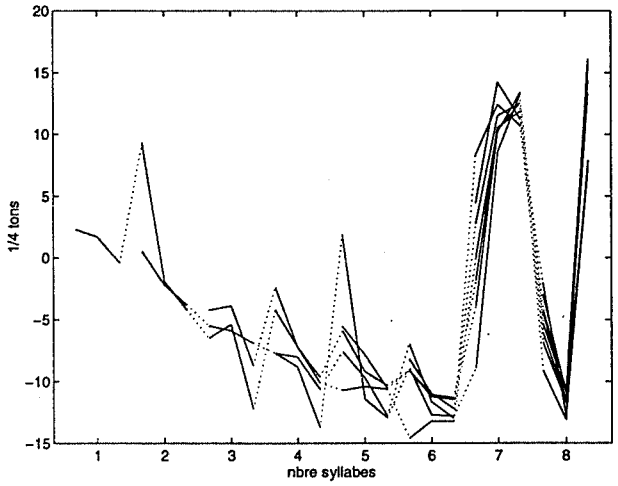


Figure 3 : Moyennes des F0 stylisées de phrases monomot de 1 à 8 syllabes pour le doute-incrédulité.

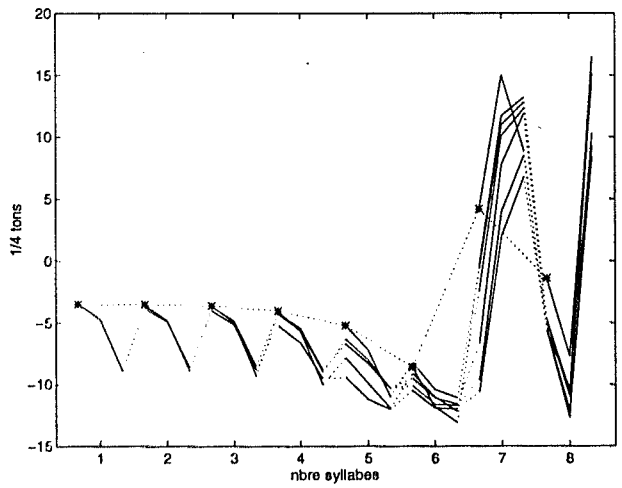


Figure 4 : Prédictions mélodiques du réseau (1 à 6 syllabes) et généralisation à 7 et 8 syllabes.

7. CONCLUSIONS

L'objectif de notre étude est double :

(a) démontrer que notre modèle morphologique de la prosodie peut émerger d'un apprentissage hiérarchique de contours. Cette focalisation de l'apprentissage sur divers niveaux linguistiques est à rapprocher de la théorie Spotlight (Strömqvist & al., 1995) : la prosodie sert à enrichir la compétence linguistique de l'enfant par « zooms » successifs.

(b) démontrer que cette approche produit des systèmes de génération de la prosodie performants et évolutifs.

Ce modèle de superposition de « gestes » prosodiques est semblable à des modèles de contrôle gestuels de la parole (la mâchoire comme organisateur syllabique est porteuse des gestes labiaux et linguaux) et des modèles d'écriture manuscrite (le bras gère le déplacement linéaire et porte les mouvements de la main formant les lettres).

8. BIBLIOGRAPHIE

Petitot-Concordat J. (1986) *Le « morphological turn » de la phénoménologie. Chapitre I, II et III de morphogénèse du sens II*, Centre d'analyse et de Mathématique Sociales, EHESS - CNRS

Thorsen N. (1980) A study of the perception of sentence intonation - evidence from Danish, *Journal of the Acoustical Society of America*, 67(3):1014-1030

Fujisaki H. and Sudo H. (1971) A generative model for the prosody of connected speech in japanese, *Annual Report of Engineering Research Institute*, 30:75-80

Aubergé V. (1992) Developing a structured lexicon for synthesis of prosody, In Bailly, G., Benoît, C., editors, *Talking Machines: Theories, Models and Design*, Elsevier B.V. 307-321

Strömqvist S., Peters A. & Ragnarsdóttir H. (1995) Particles and prepositions in Scandinavian child language development : effects of prosodic spotlight ? *International Congress of Phonetic Sciences*, Stockholm, Sweden, 4:38-45

Silverman K. (1987) *The Structure and Processing of Fundamental Frequency Contours*, Ph.D. thesis, Cambridge University, Cambridge UK

Morlec Y., Aubergé V., & Bailly G. (1995) Evaluation of automatic generation of prosody with a superposition model, *International Congress of Phonetic Sciences*, Stockholm - Sweden, 4: 224-227

Calbris G. & Montredon J. (1981) *Oh là là. Expression intonative et mimique. Livre du professeur*, CLE international

Elman J.L. (1988) *Finding structure in time*, CRL Tech. Rep. 8801, University of California, San Diego

UN SYSTÈME PRÉDICTIF DE LA STRUCTURATION SYNTAXICO-RYTHMIQUE D'UN ÉNONCÉ À L'AIDE D'INFORMATIONS PROSODIQUES.

Philippe LANGLAIS†, Jean-Luc COCHARD‡, Henri MÉLONI†

†Laboratoire d'Informatique – 339, Chemin des Meinajariès – BP 1228 – 84911 Avignon Cedex 9
Tél.: 90 84 35 25 – Fax: 90 84 35 01 – e-mail: {langlais,meloni}@univ-avignon.fr
‡IDIAP – 4 rue du Simplon – CH-1920 Martigny
Tél.: (+41) 26 22 76 64 – Fax: (+41) 26 22 78 18 – e-mail: cochard@idiap.ch

ABSTRACT

An automatic correlative system has been elaborated ; firstly, it upholds an assistance to prosodic analysis (providing vizualization and query tools), and secondly, it gives a predictive function of the linguistical structure of the message to decode. Two applications of this system are proposed ; a first one for the recognition of decimal numbers (our system is able to locate the word "virgule" in an unknown number, only by means of prosodic information) and a second one for the recognition of read isolated sentences. The results obtained fully validate the approach we proposed.

1. INTRODUCTION

Plusieurs études ont montré par le passé les nombreuses possibilités d'utilisation de la prosodie dans les systèmes de reconnaissance automatique de la parole (RAP)[Lea 80, Vaissière 88]. Les systèmes existants [Carbonell 82, Bonin 90, Nasri 90] ont cependant tous contribués, malgré eux, à montrer que l'énoncé de règles prosodiques aussi complexes soient-elles, n'est pas d'une grande efficacité en reconnaissance automatique de la parole. Les principales raisons de cet échec sont premièrement analysées introduisant ainsi notre système de prédiction automatique de la structure syntaxico-rythmique d'un énoncé. Les résultats de prédiction mesurés sur deux bases de parole continue (phrases et nombres décimaux) sont ensuite présentés.

2. POSITION DU PROBLÈME

2.1. Difficultés

Parmi les nombreuses difficultés de l'intégration de la prosodie dans les systèmes de RAP, il convient de distinguer celles qui sont imputables aux systèmes (qui soulèvent entre autre le problème fondamental et non trivial de la fusion des connaissances) de celles qui sont inhérentes à l'analyse prosodique.

D'un point de vue purement prosodique, et abstraction faite des problèmes non réglés totalement de la mesure des paramètres et de leur correction microprosodique et perceptuelle, nous rappelons la principale cause de l'échec à l'intégration de la prosodie en reconnaissance : la prosodie est le fruit d'une interaction complexe entre différents niveaux de structuration du message (syntaxique, sémantique, pragmatique, rythmique, ...) qui peuvent être conflictuels [Vaissière 88].

2.2. Quelles solutions ?

Face à la multiplicité des facteurs intervenant sur les variations des paramètres prosodiques, il semble difficile à un expert de mener à bien son analyse sans y introduire un biais que VanSanten appelle la "piecemeal analysis" (analyse locale de quelques facteurs a priori importants) [Santen 94]. Cet auteur remarque alors très justement que le recours à l'outils statistique permet de s'affranchir de cette analyse locale qui n'est pas souhaitable. On assiste d'ailleurs depuis quelques années à une tendance de plus en plus marquée au remplacement des connaissances formalisées par des experts par des composantes statistiques [Veilleux 93, Pagel 95].

Nous pensons qu'une solution satisfaisante doit tenir compte des avantages de chacune des deux approches et nous préconisons l'usage de méthodes statistiques pour épauler l'expert dans son analyse. C'est en ce sens qu'a été conçu le système que nous allons présenter.

3. LE SYSTÈME

3.1. Étiquetage prosodique

L'étiquetage prosodique consiste à réduire l'ensemble des paramètres prosodiques à un nombre limité d'indices pertinents, avec le minimum de perte d'information. Ce problème délicat n'a actuellement pas de solution unanime (cf. le workshop de Stockholm'95 sur ce thème). En l'absence de consensus, nous avons retenu un jeu assez classique de 40 étiquettes (9 indices de durée, 9 indices d'intensité et 22 indices pour la f0) qui caractérisent chaque noyau vocalique : maximum et minimum de chaque paramètre sur l'énoncé, émergence de la valeur d'un paramètre sur un noyau (n) par rapport aux valeurs correspondantes sur les segments adjacents (n-2,n-1,n+1,n+2), différents codages en niveaux (1 à 4) de la valeur de f0 d'un noyau vocalique sur une échelle correspondant au découpage en 4 zones de la dynamique du paramètre sur l'ensemble de l'énoncé, etc.

Cet étiquetage simple et entièrement automatique (nécessitant pour seule entrée le signal de parole) ne fait intervenir aucune étape de correction microprosodique ou perceptive et ce en raison des expérimentations et remarques exposées dans [Langlais 95].

Cette étape de caractérisation paramétrique est par nature criticable, aussi nous contentons nous de remarquer que notre système est ouvert à l'ajout de nouvelles étiquettes (pour autant qu'elles soient automatiquement calculables) ou au contraire à la suppression d'autres.

3.2. Le principe

Nous émettons l'hypothèse que la distribution des configurations prosodiques sur la totalité d'un énoncé n'est pas du tout aléatoire mais est au contraire suffisamment régulière pour autoriser des prédictions structurelles à partir des seules informations prosodiques.

Pour vérifier cette hypothèse, nous avons élaboré le système ProStat qui propose deux fonctionalités :

1. l'aide à l'analyse prosodique de grand corpus par un expert. Le système dispose d'une interface graphique permettant de visualiser des contours paramétriques divers. Il renseigne (par l'usage d'un jeu restreint de requètes) son utilisateur sur les corrélations (mesurées sur un corpus donné) entre des indices prosodiques et différents niveaux d'organisation ou points particuliers du message.

2. La prédiction de la structure linguistique d'un énoncé à partir des seuls indices prosodiques calculés automatiquement. Le travail de l'expert consiste uniquement à rassembler un corpus de parole et d'en fournir la description linguistique (syntaxique et/ou sémantique, etc.).

Dans cette étude, nous avons décidé de réduire notre champ d'analyse aux seules interactions du rythme (que nous définissons ici par la distribution du nombre de voyelles des différents groupes d'un énoncé) et de la syntaxe sur la distribution des indices prosodiques.

Le principe de base que nous ne développerons pas, faute de place, est la création automatique d'un graphe orienté, à partir d'un corpus d'apprentissage [Langlais 95, pp. 141 à 144]. Chaque énoncé du corpus est décrit par sa décomposition grammaticale sous forme arborescente (fournie manuellement dans notre cas), par un alignement phonétique obtenu par des modèles markoviens développés à l'IDIAP et par son treillis prosodique automatiquement calculé à partir du signal de parole. Chaque arc de ce graphe est une contrainte de nature syntaxique et/ou rythmique dérivée du corpus d'apprentissage. Chaque nœud du graphe décrit une structure syntaxico-rythmique particulière et contient des informations comme le nombre de fois où il a été visité lors de l'apprentissage, le nombre d'occurrences de chaque étiquette prosodique apposée à l'initiale ou en finale d'un groupe quelconque de la structure décrite, etc. Plus on avance dans le graphe et plus la structure syntaxico-rythmique décrite est précise.

En dotant notre système d'une métrique simple, capable de fournir une distance entre le treillis prosodique d'un énoncé inconnu et les informations contenues dans un nœud donné du graphe, nous disposons d'un outil capable de réaliser des prédictions syntaxico-rythmiques.

Nous allons maintenant décrire deux expériences qui valident notre système et confirment l'hypothèse que nous avons formulée.

3.3. Résultats

Ces deux expériences ont commencé par une étape préalable d'apprentissage (que nous décrivons brièvement pour chaque tâche) qui a nécessité l'intervention d'un expert pour fournir les décompositions syntaxiques sous forme arborescente de chaque énoncé des corpus d'apprentissage.

Nous avons ensuite demandé à notre système de formuler des hypothèses syntaxico-rythmiques valuées complètes (arbre syntaxique complet ainsi que le nombre de voyelles de chacune de ses feuilles) à l'aide du graphe issu de l'apprentissage et des treillis prosodiques automatiquement calculés pour des énoncés issus de différents corpus de tests.

3.3.1 Les phrases isolées

Un corpus de 500 phrases isolées, répétitions en nombre inégal de 80 phrases différentes par 50 locuteurs à travers un canal téléphonique assez bruité a ici servi de corpus d'apprentissage. Ces phrases de structures syntaxiques simples (principalement des phrases composées dans un ordre variable d'un groupe sujet, d'un groupe verbal et d'un ou de plusieurs compléments circonstanciels) contenaient de 4 à 17 voyelles.

ProStat a montré des capacités très intéressantes qui permettent son utilisation pour la reconnaissance de phrases :

1) sur l'ensemble des énoncés utilisés pour l'apprentissage, plus de 90% des 500 phrases sont classées en première position (parmi un choix moyen de 15 possibilités) ;

2) sur un corpus de test de 300 phrases (la majorité étant des répétitions différentes des 80 phrases de base du corpus d'apprentissage, pas nécessairement prononcées par les mêmes locuteurs), la première hypothèse affecte correctement près de 60% des énoncés à la bonne structure. Trois à cinq hypothèses suffisent pour assurer l'association de l'énoncé à sa structure syntaxico-rythmique (voir figure 1).

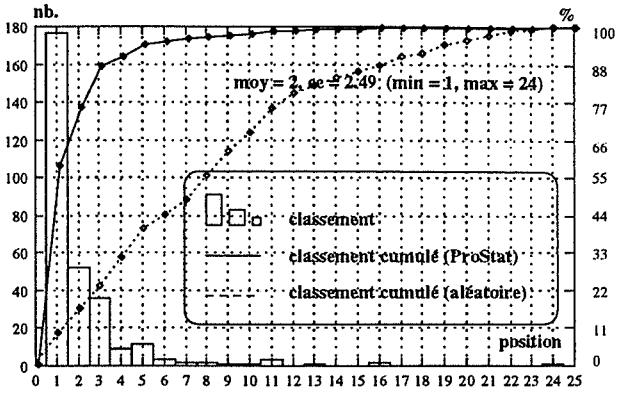


Figure 1. *Classement des hypothèses fournies par ProStat pour les phrases du corpus de test (abscisse = rang de la bonne proposition, ordonnée à gauche = nb. de propositions formulées, ordonnée à droite = pourcentage cumulé de propositions justes). La courbe en pointillé correspond à une proposition aléatoire.*

3.3.1 Les nombres décimaux

Un corpus de 500 nombres décimaux prononcés via le même canal téléphonique par une cinquantaine de locuteurs a servi de corpus d'apprentissage dans cette expérience (une grammaire classique des nombres a ici fourni les arbres grammaticaux de chaque nombre du corpus).

Là encore le système s'est avéré apte à associer les treillis prosodiques des nombres du corpus d'apprentissage à leur bonne structure syntaxico-rythmique dans plus de 80% des cas (avec en moyenne 15 structures possibles par nombre).

Sur un corpus de test de 298 nombres, le système a montré un taux de prédiction en tête d'un peu moins de 50%, ce qui tend à indiquer que l'information prosodique des nombres décimaux est moins riche que celle des phrases prononcées isolément.

En analysant les hypothèses non classées premières, il est apparu très clairement que

le mot "virgule" était très souvent bien positionné (voir figure 2). Nous n'avons pas réussi à approcher ce score par la formulation de règles locales spécifiques, ce qui semble confirmer l'importance de l'information prosodique prise sur la globalité d'un énoncé.

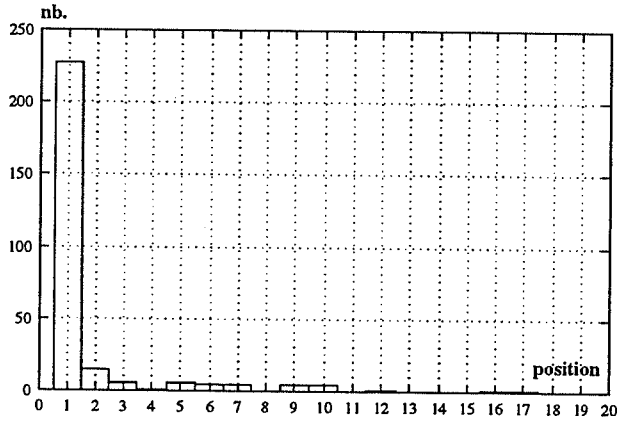


Figure 2. *Classement des hypothèses fournies par ProStat pour les nombres du corpus de test en considérant uniquement l'exactitude de la position du mot virgule dans la chaîne.*

4. CONCLUSIONS

De ces deux expériences, nous pouvons retenir que l'information prosodique — même extraite automatiquement — permet de formuler des hypothèses syntaxico-rythmiques avec un taux de réussite significatif. Le choix de notre approche sans connaissance *a priori* permet également de valider l'hypothèse d'une distribution non aléatoire des indices prosodiques sur la globalité d'un énoncé.

Si l'on peut à nouveau convenir que l'information prosodique doit être intégrée avec profit dans les systèmes de reconnaissance de la parole, nous n'avons cependant pas montré, pour chaque type d'utilisation, les efforts importants qui restent nécessaires pour aboutir à un emploi optimal de ces données et de ces connaissances. Des travaux supplémentaires doivent être menés afin de définir — pour une tâche donnée — un sous-ensemble d'indices prosodiques pertinents (actuellement, chaque indice participe de manière égale à la notation par le système des différentes hypothèses structurelles) ; les tests doivent de plus être étendus à des énoncés aux structures plus variées.

Bibliographie

[Bonin 90] J.J. Bonin et J.M. Pierrel. Fréquence fondamentale et durée pour la détection de frontières syntagmatiques en parole continue. Dans *XVIIIème JEP*, Montréal, 1990.

[Carbonell 82] N. Carbonell, J.P. Haton, F. Lonchamp, et J.M. Pierrel. Élaboration expérimentale d'indices prosodiques pour la reconnaissance ; application à l'analyse syntaxico-sémantique dans le système Myrtille II. Séminaire Prosodie et Reconnaissance d'Aix-en-Provence, Octobre 1982.

[Langlais 95] P. Langlais. *Utilisation de la prosodie en reconnaissance automatique de la parole.* Thèse, Université d'Avignon, 1995.

[Lea 80] Wayne A. Lea. Prosodic aids to speech recognition. Dans *Trends in Speech Recognition*, W.A. Lea, éditeur. Prentice Hall, 1980.

[Nasri 90] M.K. Nasri, G. Caelen-Haumont, et J. Caelen. Utilisation de règles prosodiques en reconnaissance de la parole. Dans *XVIIIèmes JEP*, 1990.

[Pagel 95] V. Pagel, N. Carbonell, et J. Vaissière. Spotting prosodic boundaries in continuous speech in french. Dans *XIIIth International Congress of Phonetic Sciences*, volume 4, pages 308–311, Stockholm, 1995.

[Santen 94] J.P.H. van Santen. Using statistics in text-to-speech system construction. Dans *Second ESCA/IEEE workshop on Speech Synthesis*, pages 240–243, New York, 1994.

[Vaissière 88] J. Vaissière. The use of prosodic parameters in automatic speech recognition. Dans *Recent Advances in Speech Understanding and Dialog Systems*, volume F46. NATO ASI Series, 1988.

[Veilleux 93] N.M. Veilleux et M. Ostendorf. Probabilistic parse scoring with prosodic information. Dans *IEEE International Conference on Acoustics, Speech and Signal Processing*, volume II, pages 51–54, 1993.

VARIATION DE DÉBIT NASAL EN FONCTION DE LA POSITION PROSODIQUE DE [n] ET [ã] EN FRANÇAIS.

Cécile Fougeron

Lab. de Phonétique, Paris III - URA 1027 & UCLA, e-mail : fougeron@humnet.ucla.edu

ABSTRACT

Nasal airflow for [n] and [ã] in French is found to vary depending on the prosodic position of the segment. Although the positions that may be distinguished by the amount of nasal flow vary between speakers, segments occurring in higher prosodic positions have less nasal flow than those occurring in lower positions. This prosodic effect is more consistent for [n] than for [ã]. The consonants in higher prosodic positions are also more distinct from the following vocalic nucleus in term of acoustic energy.

1- INTRODUCTION.

L'étude des relations entre prosodie et articulation soulève deux questions : l'articulation d'un segment varie-t-elle en fonction de sa position prosodique ? Cette variation reflète-t-elle la hiérarchie des constituants prosodiques ?

La parole est organisée selon une structure prosodique qui peut être définie en terme de constituants prosodiques [Selkirk 1984, Nespor & Vogel 1986]. Ces constituants sont sont organisés de manière hiérarchique et définissent le domaine d'application de certaines règles phonologiques. Par exemple, le Mot Phonologique est dominé par la Phrase Phonologique, elle-même dominée par la Phrase Intonative. La "position prosodique" signifie la position d'un segment dans l'un de ces constituants.

Notre étude examine comment la position d'un segment dans la structure prosodique peut influencer son articulation. Il a été récemment montré que l'articulation glottale de certains segments est renforcée lorsque ces segments se situent au début d'un constituant prosodique de "haut niveau". En anglais, le Voice Onset Time (V.O.T.) de [t] et le bruit d'aspiration de [h] sont plus longs lorsque les segments débutent une Phrase Intonative que lorsqu'ils sont au milieu de cette phrase (Pierrehumbert & Talkin 1992). En Coréen, le V.O.T. de [p^h] est plus long au début d'une Phrase Accentuelle qu'au début d'un Mot Lexical, et il est moins long au milieu d'un Mot Lexical qu'au début d'un Mot (Jun 1993). L'articulation linguale varie aussi en fonction de la position prosodique [Fougeron et Keating 1995]. Lors de l'articulation de la consonne [n] en anglais, l'aire de contact de la langue sur le palais est plus large au début qu'elle ne l'est au milieu ou à la fin d'un constituant prosodique. De plus, l'aire de contact diminue progressivement en fonction de la hauteur du constituant dans la hiérarchie prosodique. Les résultats montrent qu'un [n] situé au début du constituant le plus élevé (Phrase Intonative) a plus de contact linguopalatal qu'un [n] situé au début d'un constituant de niveau hiérarchique intermédiaire (Phrase Phonologique). Ce dernier a, à son tour, plus de contact qu'un [n] au début d'un constituant de bas niveau (Mot Lexical). En résumé, il semblerait qu'il existe un "renforcement" des articulations glottale et orale au début des constituants prosodiques de haut niveau hiérarchique, c'est à dire lorsque le segment est à une position "forte" du point de vue prosodique.

L'objet de cette étude est d'évaluer l'influence de la position prosodique sur l'articulation vélaire d'un segment nasal en français. Si on extrapole les résultats présentés ci-dessus pour prédire l'effet de la position prosodique sur l'articulation nasale, on pourrait s'attendre à ce que le renforcement articulatoire en position prosodique "forte" se traduise par une aperture vélopharyngale plus large. Ainsi, un segment nasal en position forte aura un débit d'air nasal plus important, et sera donc plus nasal.

Cette prédiction n'est pas vérifiée dans le cas des variations de l'articulation vélaire des consonnes nasales en fonction de leur position dans le mot ou la syllabe (i.e. des constituants de très bas niveau). Il a été observé que l'aperture vélopharyngale est plus étroite lorsque les consonnes nasales sont placées en début plutôt qu'en fin de mot ou de syllabe (Benguerel 1977 pour le français, Fujimura 1977 pour l'anglais et le japonais, Krakow 1989 pour l'anglais). Par conséquent, les consonnes nasales en début de mot ou de syllabe sont moins sonores et donc plus consonantiques (Manuel 1991).

A la lumière de ces observations, il est possible d'émettre l'hypothèse que l'articulation nasale va varier, tout comme l'articulation glottale ou orale, en fonction de la position prosodique des segments : plus la position prosodique du segment sera forte, plus son aperture vélopharyngale sera petite.

En d'autres termes, lorsqu'un segment est à une frontière prosodique importante, on prédit que son débit nasal sera plus faible.

2- MÉTHODE

Deux segments nasals ont été étudiés : la consonne nasale coronale [n] et la voyelle nasale ouverte [ɑ̃]. La consonne nasale [n] est examinée dans deux types de contexte vocalique variant en fonction de l'impédance orale de la voyelle : forte pour [i_i] et faible pour [a_a].

Les segments tests sont placés dans quatre positions prosodiques différentes. Le tableau 1 donne un exemple des stimuli formés avec la séquence [ana]. Dans le stimulus 1, la consonne [n] est en début de Phrase ("Utterance", Ui). Elle est séparée de la Phrase précédente par une longue pause. Dans le stimulus 2, [n] est placé au début d'une Phrase Intonative (PIi). Ce constituant est délimité à droite par un fort allongement prépausal et une longue pause. Dans le stimulus 3, [n] se trouve au début d'une Phrase Accentuelle (PAi). La Phrase Accentuelle est le plus bas des constituants prosodiques définis par l'intonation en français [Jun & Fougeron, 1995]. Elle est caractérisée par un allongement final moyen et une montée de continuation mineure. Dans le stimulus 4, [n] est placé au milieu de la Phrase Accentuelle "Tata Nadia" et au début du mot "Nadia" (PAm). Il n'est pas certain que le Mot Lexical soit un constituant prosodique, on considérera donc cette position comme PAm.

Des stimuli similaires ont été construits pour la séquence [i_i] avec la suite "Tatie Nicole". Pour la voyelle nasale [ɑ̃], les stimuli ont été formés avec la suite "Jacques André", et la voyelle a été placée à 3 positions prosodiques : au début d'une Phrase Intonative, au début d'une Phrase Accentuelle, et au milieu d'une Phrase Accentuelle.

4 locuteurs français parisiens (3 femmes et 1 homme) ont été enregistrés en deux sessions comprenant 10 répétitions de chaque phrase pour le corpus avec [n] et en une session de 10 répétitions pour le corpus avec [ɑ̃]. Les locuteurs n'ont pas reçu d'instruction quant à la façon de lire ou de "phraser" les stimuli. Lors de l'analyse, les énoncés qui ne correspondaient pas à la structure prosodique prévue dans les stimuli ont été éliminés.

Les débits d'air nasal et oral ont été enregistrés à l'aide d'un masque de type Rothenberg et analysés avec le système CSL-Kay. Les variations de débit d'air nasal sont considérées comme une mesure indirecte des modifications de l'aperture vélopharyngale. Différentes mesures (spatiales et temporelles) ont été prises sur l'onde de débit nasal, mais dans cette article il ne sera question que du maximum (ou "pic") de débit nasal dans la consonne [n] et la voyelle [ɑ̃].

Du point de vue acoustique, les variations de débit d'air nasal ont été interprétées en terme d'énergie acoustique. Pour une consonne nasale, la quantité d'air s'échappant par le nez se traduit par une variation d'énergie acoustique. En revanche, l'énergie d'une voyelle nasale dépend aussi bien du débit nasal que du débit oral. Par conséquent, les maxima d'énergie acoustique (calculés à partir du signal audio) ont été relevés seulement pour les consonnes. Comme les niveaux d'énergie varient d'une répétition à l'autre, les comparaisons se font sur des mesures d'énergie relatives au maximum d'énergie dans la voyelle suivant le [n].

Tableau 1 : Stimuli formés avec la séquence [a**n**a]. Les abréviations utilisées dans les graphes suivants sont indiquées entre parenthèses (voir le texte).

1- Début de Phrase (Ui) : "Paul aimait Tata. **N**adia, par contre, aimait Paul."
2- Début de Phrase Intonative : (PIi) "Pauvre Tata, **N**adia et Paul n'arriveront que demain."
3- Début de Phrase Accentuelle : (PAi) "Tonton, Tata, **N**adia et Paul arriveront demain par le train."
4- Milieu de Phrase Accentuelle : (PAm) "Tonton-Paul et Tata-**N**adia arriveront demain par le train."

3. RÉSULTATS : CONSONNE [n]

3.1. Maximum de débit nasal pour [n]

Les maxima de débit d'air nasal varient en fonction de la position prosodique de la consonne [n] ($F=13.3$, $p<.001$). Lorsqu'on regroupe les résultats des 4 locuteurs, on observe une diminution progressive de débit au fur et à mesure que la position de [n] s'élève dans la hiérarchie prosodique. Les [n] placés au milieu d'une PA ont un débit d'air nasal plus important que ceux placés au début d'une PA. Les [n] situés au début d'une PI ont à leur tour un débit nasal plus important que ceux placés au début du constituant inférieur, la PA. Par contre il n'y a pas de distinction entre les positions PIi et Ui. En résumé, trois positions prosodiques sur quatre sont marqués par une différence de débit nasal.

Ceci dit, l'effet de la position prosodique de [n] sur son débit nasal varie en fonction du locuteur (interaction $F=9.1$, $p<.001$). Cette variation porte principalement sur le nombre et la nature des positions prosodiques distinguées par une différence de débit nasal. La figure 1 présente les maxima de débit nasal pour les [n] placés dans les 4 positions prosodiques

considérées pour chaque locuteur.

Le locuteur 1F ne distingue pas les positions PAm et PAi. Par contre, les [n] placés au début d'une PI ont un débit nasal significativement plus faibles que ceux placés au début dune PA (t=8, p<.001) ou au milieu d'une PA (t=6, p<.001). Les positions PIi et Ui ne sont pas distinguées par le débit nasal.

Pour le locuteur 2F, la distinction entre PAm et PAi n'est significative que pour le contexte vocalique [ana] (t=2, p=.04). Pour les deux contextes vocaliques, les [n] placés au début d'une PI ont un débit nasal plus faibles que ceux placés au début dune PA (t=5, p<.001) ou au milieu d'une PA (t=7, p<.001). De même, les [n] en position Ui ont un débit plus faible qu'au début de PI (t=4, p<.001).

Le locuteur 3M distingue les positions PAm et PAi (t=2, p=.03). Les [n] en début de PI ont un débit plus faible qu'au début de PA (t=4, p<.001) ou au milieu de PA (t=7, p<.001). Pour cette comparaison avec la position PIi, la 1ère session d'enregistrement de ce locuteur est exclue. Dans cette session, le débit des [n] en début de PI est rehaussé par la superposition de l'expiration pendant la pause avec le début de la consonne. Les [n] en début d'U ont un débit plus fort que dans toutes les autres positions.

Le locuteur 4F ne distingue aucune position prosodique par une variation significative de débit nasal. Les [n] au début de PA tendent à avoir un débit plus faible qu'au milieu de PA, mais aussi plus faible qu'au début de PI. Après examen du phrasé produit par ce locuteur, il s'avère qu'il ne distingue, ni par l'intonation ni par la pause, les stimuli de type 2 (PIi) et 3 (PAi).

En résumé, 3 des 4 locuteurs distinguent au moins deux positions prosodiques mais varient dans la distinctions des autres positions :

(PAm & PAi) > (PIi & Ui)	pour le loc. 1F
(PAm & PAi) > PIi > Ui	pour le loc. 2F
PAm > PAi > PIi < Ui	pour le loc. 3M

3.2. Différences d'énergie acoustique.

Les conséquences acoustiques d'une variation de débit nasal pour une consonne se traduisent directement par une variation d'énergie acoustique. La figure 2 présente les différences de maxima d'énergie acoustique de [n] par rapport à ceux de la voyelle suivante, pour les trois locuteurs marquant une distinction entre les différentes positions prosodiques.

Plus la position de [n] monte dans la hiérarchie prosodique, plus la consonne se distingue de la voyelle suivante en terme d'énergie acoustique. Les différences en décibel sont faibles mais sont significatives pour les 3 positions (PAm, PAi et PIi), sauf pour les positions PAm et PAi du locuteur 2F. La corrélation entre les maxima de débit nasal et les différences d'énergie acoustique est très faible (r^2=.1 tous loc. confondus). Pourtant, si l'on compare les figures 1 et 2, il apparaît clairement que les différences d'énergie et le débit nasal sont en relation inverse : moins le débit nasal est fort, plus la distinction énergétique entre [n] et V2 est grande.

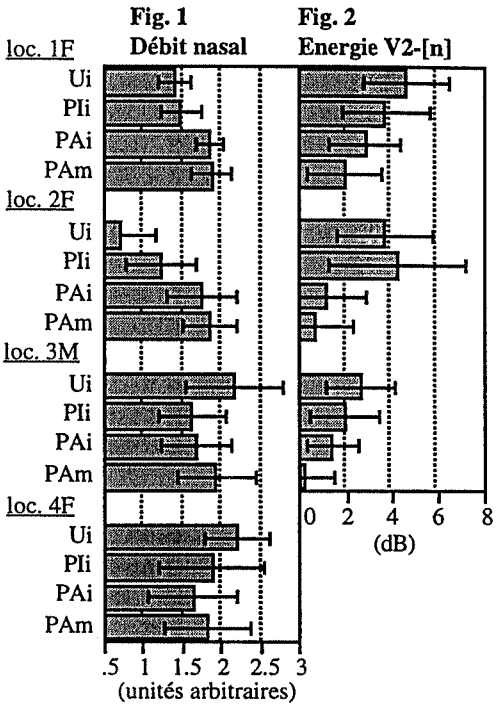


Figure 1 et 2 : 1. Maxima de débit d'air nasal dans la consonne [n] en fonction de sa position prosodique et par locuteur. 2. Différence entre les maxima d'énergie acoustique de [n] et de la voyelle suivante.

4- RÉSULTATS : VOYELLE [ã]

La figure 3 présente les maxima de débit d'air nasal pour [ã] dans les 3 positions prosodiques examinées pour chaque locuteur. Le loc. 1F ne marque aucune distinction entre les différentes positions prosodiques pour la voyelle [ã] alors qu'il distinguait deux niveaux prosodiques pour la consonne [n]. Le locuteur 2F distingue de manière significative les trois positions avec une diminution progressive du débit nasal de la position la plus basse à la position la plus haute dans la hiérarchie prosodique. Pour le locuteur 3M, les [ã] placés au milieu d'une PA ont un débit plus important que ceux placés au début d'une PA (t=4, p=.003). La distinction entre les positions PAi et PIi ou PAm et PIi n'est pas significative. Il a été noté précédemment que la position PIi est souvent problématique pour ce locuteur qui expire tard dans la pause. Le locuteur 4F qui ne marquait aucune distinction pour la consonne [n], ne distingue que les positions PAi et PIi. (t=3, p=.008) sans différencier les positions PAm et PIi.

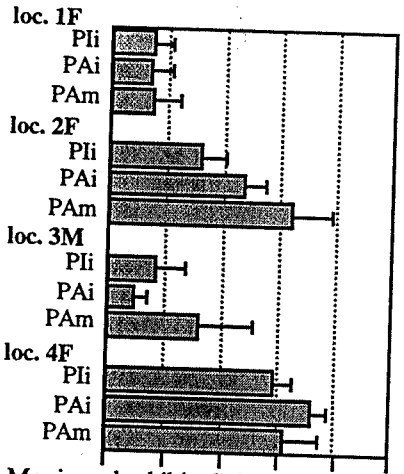


Figure 3 : Maxima de débit d'air nasal dans la voyelle [ɑ̃] en fonction de sa position prosodique.

5. DISCUSSION ET CONCLUSION:

L'articulation nasale de [n] est influencée par sa position prosodique pour 3 locuteurs sur 4. La variation articulatoire observée reflète la hiérarchie prosodique des constituants : le débit d'air nasal diminue au fur et à mesure que la position du segment s'élève dans cette hiérarchie. En d'autres termes, lorsqu'un segment est à une frontière prosodique importante sont débit nasal est plus faible. Cette variation d'articulation vélaire ne permet pas de distinguer tous les niveaux prosodiques examinés. Ces 3 locuteurs distinguent au moins 2 niveaux prosodiques : un niveau "inférieur" (PAm, PAi) et un niveau "supérieur" (PIi, Ui). Deux des trois locuteurs marquent une distinction supplémentaire à l'intérieur d'un de ces sous-groupes. Le début de Phrase (Ui) est la position la plus variable. Son statut prosodique n'est donc pas résolu. Cette technique d'investigation aérodynamique n'est d'ailleurs peut être pas appropriée pour l'examen de ce constituant qui peut former un nouveau groupe de souffle avec un débit d'air total qui n'est pas comparable aux autres cas.

Pour la voyelle nasale [ɑ̃], les résultats sont moins clairs. Seul 1 locuteur sur 4 distingue les trois positions prosodiques examinées par une variation de débit nasal. Un des locuteurs ne marque aucune différence et les deux autres locuteurs font des distinctions isolées entre deux différents niveaux prosodiques contigus. Lorsqu'une variation articulatoire est observée, la tendance est la même que pour la consonne [n] : plus la position est haute dans la hiérarchie prosodique, plus le débit nasal est faible.

Les résultats obtenus avec d'autres articulateurs pourraient suggérer que le renforcement articulatoire pour les positions prosodiques "fortes" est une forme d'hyperarticulation locale : l'amplitude de l'occlusion orale des consonnes est plus importante (Fougeron & Keating 1995) et la position du vélum est plus élevé pour les consonnes orales (Vaissière 1988). Nos résultats ne corroborent pas cette hypothèse pour l'articulation nasale des segments nasals pour lesquels le renforcement articulatoire est marqué par une *réduction* de l'aperture vélopharyngale aux positions prosodiquement fortes. Cette réduction n'augmente donc pas la nasalité du segment tout au contraire. Il semblerait donc peut être plus approprié de considérer cette variation articulatoire en terme de "saillance" acoustique ou perceptuelle. L'examen des différences d'énergie acoustique suggère que plus la position du segment est forte plus la consonne est distincte, en terme d'énergie, du nucleus vocalique suivant. Ces différences d'énergie permettent d'ailleurs de distinguer un nombre plus important de positions prosodiques que le débit nasal seul. Ces différences sont faibles en décibel et sont comparable aux JND d'énergie pour un ton pur (.3 dB à 80 dBSL) ou d'un bruit à bande large (.5 à 1 dB) (Moore 1982). Toutefois, il n'est pas certain que ces différences de maxima d'énergie soient perceptibles ni qu'elles soient utilisées dans le processus de perception.

Ce travail est financé en partie par une allocation du M.R.T. attribuée au D.E.A. de Phonétique de Paris et par la bourse NSF SBR-9511118 attribuée à P. Keating.

RÉFÉRENCES :

Benguerel, A.-P. (1977), Velar coarticulation in French: a fiberscopic study. *J.o.P.* 5:149-158

Fougeron, C. & P. Keating (1995) Demarcating prosodic groups with articulation. *J.A.S.A.* 97-5, pt.2

Fujimura, O. (1977), Recent findings on articulatory processes . In *Articulatory modeling and phonetics* (Carré, Descout & Wajskop eds.), GALF, 115-126.

Jun, S.-A. (1993), *The phonetics and phonology of Korean prosody*. PhD diss., Ohio State U.

Jun, S.-A. & C. Fougeron (1995), The Accentual Phrase and the prosodic structure of French. *ICPhS 95*, 722-25

Krakow, R. A. (1989), The articulatory organization of syllables: a kinematic analysis of labial and velic gestures. PhD. diss.Yale U.

Manuel, S. (1991), Some phonetic bases for the relative malleability of syllable-final vs. syll.-initial consonants. *ICPhS 91*, 5:118-121

Moore, B. (1982), *An introduction to the psychology of hearing*. Academic Press.

Nespor, M. & I. Vogel (1986), *Prosodic Phonology*. Dodrecht: Foris Publications.

Pierrehumbert, J. & D. Talkin (1992), Lenition of /h/ and glottal stop. In *Lab. Phonology II* (Docherty & Ladd eds.) Cambridge U. Press.

Selkirk, E. (1984), *Phonology and syntax: the relation between sound and structure.* MIT Press.

Vaissière, J. (1988), Prediction of velum movement from phonological specifications. *Phonetica* 45:122-39.

VERS UNE TYPOLOGIE DES UNITÉS INTONATIVES DU FRANÇAIS.

Albert DI CRISTO et Daniel HIRST

Institut de Phonétique d'Aix - URA CNRS 261 .Parole et Langage,
Université de Provence - 29 avenue Schuman - 13621 Aix-en-Provence
Tél.: +33 42 953618 - Fax: +33 42 595096 - E-mail dicristo@lpl.univ-aix.fr

ABSTRACT

In this paper we present a preliminary typology of a number of intonation patterns which are constitutive of some basic Intonation Units (UI) in French. In the first part we introduce the theoretical framework of our parametric approach. In the second part we show that these patterns can be derived from a phonological model involving underlying High and Low tones and well-attested principles of linearisation, detachment, lowering and raising.

1. INTRODUCTION.

Des études récentes relatives à la production et à la perception de la parole (Beckman & Edwards, 1992; Goodman & al., 1994) tendent à corroborer l'hypothèse linguistique suivant laquelle la fonction essentielle de la prosodie est *d'organiser la forme sonore du langage*. Dans le cadre de la "phonologie prosodique" (autosegmentale et métrique), la prise en compte de ce rôle majeur de la prosodie a conduit à l'élaboration de la *"théorie des domaines"* (Selkirk, 1978; Nespor & Vogel, 1986) qui envisage la représentation mentale du système prosodique d'une langue sous la forme d'une *structure constituante hiérarchique*. Cette théorie, ainsi que la problématique plus générale du "phrasing" qui s'y rattache, continuent de soulever de nombreux problèmes qui concernent à la fois l'inventaire des constituants nécessaires pour décrire la prosodie d'une langue, les critères qui président à leur identification, la représentation formelle de ces unités et la nature des relations entre les domaines de la structure prosodique et ceux qui relèvent des autres composantes de la grammaire comme la syntaxe et la sémantique. Bien que nous n'ayons pas l'intention de débattre ici ces questions au demeurant essentielles, le thème de notre communication relève pleinement de cette problématique . Nous nous proposons en effet d'esquisser dans ce travail une *typologie des unités intonatives de base du français*, fondée sur la mise en oeuvre des critères qui définissent le cadre théorique et méthodologique de l'approche prosodique que nous développons à l'Institut de Phonétique d'Aix depuis plusieurs années. Nous rappelons dans la première partie les grandes lignes de cette approche. Nous présentons ensuite les patrons des unités intonatives qui forment la base de notre typologie et nous illustrons leurs principaux usages.

2. FONDEMENTS THÉORIQUES

Le courant de la phonologie prosodique (métrique et autosegmentale) adopte deux types fondamentaux de représentations formelles pour décrire la structure rythmique et l'organisation intonative des langues: respectivement, *un arbre (ou une grille) métrique* et des *séquences de segments tonals*. C'est ainsi que dans le modèle intonatif désormais classique de Pierrehumbert (1980), les patrons mélodiques de l'anglo-américain sont représentés par des chaînes de segments tonals (Low) et (High) accompagnés parfois de signes diacritiques (*, %) afin de signaler des emplois particuliers (accent, frontière). Ce mode de représentation de l'intonation procède d'une option théorique fondamentale d'après laquelle il est préférable de décrire l'intonation en termes de *suites d'éléments discrets* plutôt qu'au moyen de configurations caractérisables par leurs formes globales et leurs directions.

L'approche de type paramétrique de l'intonation du français que nous proposons de développer se rallie à ce point de vue. A la différence de la théorie autosegmentale standard, elle propose cependant d'associer la ligne des segments phonémiques et celle des segments tonals par la médiation d'une structure prosodique hiérarchique comprenant à l'origine (Hirst & Di Cristo, 1984) deux couches de constituants: les Unités Tonales (UT) et les Unités Intonatives (UI). Par la suite (Di Cristo & Hirst, 1992,), afin de rendre compte de la base rythmique du français et de la distinction entre l'accent primaire (final) et l'accent secondaire (initial), nous avons proposé d'introduire dans le modèle un constituant prosodique de rang intermédiaire: l'Unité Rythmique (UR) dont il ne sera pas question ici dans la mesure ou nous traitons

uniquement de l'organisation intonative (ou mélodique) du français.

L'Unité Intonative, telle que nous la concevons, se caractérise, d'une part, par une chaîne de segments tonals constituant un *patron mélodique cohérent* et, d'autre part, par la présence obligatoire de tons spécifiques, appelés *"tons de frontière"*. La cohésion du patron mélodique de l'UI est assurée par la présence d'une configuration soit régulièrement descendante (qui illustre notamment un *phénomène global d'abaissement* attesté dans la plupart des langues: Vaissière 1995), soit régulièrement montante, soit uniformément plate. Dans cette optique, un ton indicateur de la frontière droite d'une UI peut être identifié comme tel sur la base de deux faits majeurs: d'une part, il rompt la cohésion mélodique, puisque dans une séquence régulièrement descendante de tons H, l'occurrence d'un H supérieur au précédent induit la présence d'une frontière ou d'une marque d'emphase. De plus, ce ton de frontière est associé à un schéma mélodique spécifique que nous appellerons ici *cadence*. La cadence est caractérisée par un abaissement local (Downstep) ou par un relèvement local (Upstep), suivi d'une configuration mélodique particulière: le *contour intonatif* qui constitue l'élément irréductible de l'UI. La frontière gauche de l'UI comporte, dans notre système noyau, un ton L sous-jacent qui se manifeste en surface par un effet de réajustement (resetting) au registre tonal de référence (M). Ainsi conceptualisées, les UI peuvent être analysées en termes de caractéristiques *locales*, *globales* et *itératives*, ces dernières étant représentées par le schème récurrent que décrit la succession des Unités Tonales au sein de l'UI.

Afin d'accéder au statut d'UI et recevoir ainsi un parenthésage de type [...], une séquence mélodique doit être bornée à sa droite par un ton de frontière L ou H. Si cette condition n'est pas remplie, la séquence en question constitue alors ce que nous appelons un *segment d'UI*. Les segments d'UI correspondent, soit à des UI "avortées", (fréquentes dans la parole spontanée, par suite de phénomènes d'interruption et de reprise), soit à des constructions impliquant des relations de dépendance particulières, comme on le verra par la suite. Nous considérons que le segment d'UI est rattaché à une UI "à part entière" avec laquelle il forme une *Macro Unité Intonative*, ce qui donne lieu à des parenthésages du type [...[...] ou [...]...], dans lesquels la séquence délimitée par [...[ou par]...] représente le segment d'UI

Selon notre conception, une théorie complète de l'intonation se doit, d'une part, de rendre compte de tous les *niveaux de représentation*, du plus concret au plus abstrait et, d'autre part, d'expliciter la nature de leurs interfaces. Dans cette perspective, nous postulons les quatre niveaux d'analyse suivants: *physique*, *phonétique*, *phonologie de surface* et *phonologie profonde*. La relation entre ces niveaux est fondée sur la mise en oeuvre d'un *principe d'interprétabilité* qui stipule qu'un niveau de représentation donné doit être interprétable à la fois au niveau supérieur et au niveau inférieur. C'est ainsi que la représentation physique de l'intonation, actualisée par une courbe de (F0), est interprétée au niveau supérieur: le *niveau phonétique* sous la forme d'une courbe lisse et continue constituée d'une séquence de points-cibles définissant une fonction spline quadratique (Hirst, 1980). Ces cibles sont interprétables à leur tour, au niveau supérieur ou niveau de la *représentation phonologique superficielle*, comme des événements discrets dont le codage symbolique est effectué à l'aide du système INTSINT (*INternational Transcription System for INTonation*) que nous considérons comme l'équivalent sur le plan intonatif de l'Alphabet Phonétique International.

Les symboles utilisés dans INTSINT sont des autosegments alignables avec la transcription orthographique ou phonétique d'un énoncé. Outre les crochets qui servent à parenthéser les séquences en UI et segments d'UI, les symbole relèvent de deux catégories: les *symboles absolus* correspondants à la tessiture du sujet: Mid (⇒), Top (⇑), Bottom (⇓) et les *symboles relatifs*, qui indiquent les changements de hauteur par rapport au symbole précédent: Higher (↑), Lower (↓), Same (→), Downstep (>) et Upstep (<). Une suite parenthésée de symboles de ce type constitue la représentation superficielle d'un patron intonatif, dérivable d'une structure sous-jacente (profonde) par l'application d'un ensemble de règles et de principes associés à la mise en oeuvre de divers paramètres.

3.- TYPOLOGIE DES UI DE BASE DU FRANÇAIS.

Nous formons l'hypothèse qu'au niveau sous-jacent profond l'intonation est représentée sous la forme de *schèmes (ou gabarits) tonals* qui concourent à l'identification des unités phonologiques de la langue considérée, en l'occurrence, les UI et les UT. Dans cette perspective, nous postulons que l'UI est représentée en français par deux schèmes de base: [LL] et [LH] et l'UT, par un schème

unique: **(LH)** (nous utilisons ici les symboles INTSINT, plutôt que leurs équivalents français (**B**, **H**, etc.). C'est ainsi qu'une **UI [LL]** formée de deux **UT** sera représentée par: la séquence **[L(LH)(LH)L]**. Pour qu'un énoncé correspondant à une structure tonale hiérarchique de ce type (par ex. "Il arrive à huit heures") soit prononçable, il est nécessaire d'opérer une *linéarisation* des segments tonals associés à l'**UI** et aux **UT**. Cette linéarisation découle de l'application de certaines règles prosodiques (de simplification, de faille tonale ou de downstep, etc.), qui s'inspirent des travaux relatifs à l'analyse phonologique des langues à tons (Leben, 1973; Clements & Ford, 1979). L'application ordonnée de ces règles permet de dériver, à partir des représentations de la structure profonde décrites ci-dessus, les structures de surface qui actualisent les énoncés intonativement bien formés attestables en français. Il n'est évidemment pas question de donner ici la liste complète de ces règles ni d'établir un inventaire exhaustif des **UI** de cette langue.Nous nous bornerons donc à présenter une typologie élémentaire.

3.1. UI à part entière.

Soit l'énoncé *"Une bonne bouteille de champagne"*, qui peut être la réponse à une question comme "Qu'est-ce qu'on va lui offrir pour sa fête?" Selon notre modèle, la représentation intonative sous-jacente de cet énoncé (correspondant à une seule **UI**) comporte trois unités tonales, soit trois schèmes **LH** consécutifs.

- *Pour la variante assertive* de cet énoncé correspondant au patron tonal sous-jacent **[L(LH)(LH)(LH)L]**, nous retenons deux possibilités, selon qu'il s'agit d'une réalisation non-emphatique ou emphatique. Dans le premier cas, nous identifions les deux patrons intonatifs suivants, dérivés de la représentation sous-jacente:

(a) **[MH LH D B]** (Figure 1, a)

(b) **[MH LH LHB]** (Figure 1, b)

Conformément au principe de simplification, les tons **L** initiaux de la représentation sous-jacente fusionnent en un ton moyen **M** caractéristique du registre de référence du locuteur. D'autre part, le ton **L** final se réalise comme un ton bas absolu **B** correspondant à la valeur plancher de sa dynamique tonale. Pour rendre compte de la dérivation du patron (a), qui constitue la forme la plus neutre, nous supposons le *détachement* du ton **L** de la dernière unité tonale. Comme cela a été suggéré pour les langues à tons (Clements & Ford, 1979), ce ton bas n'est pas réalisé phonétiquement, mais a pour effet d'abaisser le ton haut qui suit et qui devient de ce fait un ton abaissé (downstep) codé **D**. Dans le patron (b), qui traduit un engagement énonciatif plus marqué du locuteur et un tour plus convivial, ce détachement n'est pas effectif, ce qui confère à la cadence un schéma montant-descendant **[LHB]** au lieu du patron neutre **[DB]**.

L'emphase contrastive: "Une bonne bouteille de champagne!" ("et non une bouteille de Sauternes") engendre la réalisation en surface du patron intonatif:

(c) **[MH LH LSTB]** Figure 1, c

qui équivaut à l'intonème d'implication décrit par Delattre. L'obtention de ce patron résulte selon nous d'un processus dérivatif complexe comprenant l'assignation de tons **LH** au constituant emphatique enchâssé dans la dernière Unité Tonale **LH**, la simplification de la séquence **LLHH** ainsi formée par le détachement des tons **L** et la réalisation d'un **T** par suite du détachement de l'un des deux **H** (effet de réhaussement).

- *La variante assertive non-terminale (continuative) et la question totale* reçoivent pareillement en structure profonde les tons de frontière **[LH]**, mais ils peuvent donner lieu en surface à deux patrons mélodiques différents:

(d) **[MH LH DT]** (Figure 1, d)

et (e) **[MH D DT]** (Figure 1, e)

qui s'expliquent, dans le premier cas par le détachement du ton **L** de la dernière **UT** comme dans (a) et dans le second, par le détachement des tons **L** de toutes les **UT**, à l'exception de la première, ce qui se traduit par le caractère itératif de l'abaissement.

3.2. Segments d'UI.

Les segments d'UI correspondent à divers patrons mélodiques dont nous ne présenterons ici, faute de place, que deux types relativement fréquents. Soit les énoncés:

E1: "*Une bonne bouteille de champagne*, ça lui plairait? (question)

et E2: "Ca lui ferait plaisir, *une bonne bouteille de champagne*". (assertion)

dans lesquels les éléments dits extraposés (en italiques) constituent des segments d'**UI** d'après nos critères. En effet les représentations intonatives de surface qui les caractérisent respectivement, selon nos observations sont les suivantes:

(f) **[MH D D [MH D DH]** (Figure 1, f)

(g) **[MH LH B] B]** (Figure 1, g)

Dans (f), le segment d'UI en position initiale (support de la question) est privé de ton de

frontière. On remarque cependant qu'il présente un schéma similaire à (e), ce qui incite à penser que ce patron est commun aux énoncés interrogatifs (Di Cristo & Hirst, 1993). Dans (g), le segment d'UI en position finale (support de l'assertion) est dépourvu de ton spécifique : il se résout à une mélodie uniformément plate, qui est la copie du ton de frontière de l'UI à laquelle il est associé.

4. CONCLUSION.

La démarche que nous avons adoptée nous a conduit à poser les fondements d'une typologie formelle des unités intonatives du français. Si ce travail s'appuie sur des principes phonologiques connus, il ne perd pas pour autant le contact avec la réalité de l'observation dans la mesure où les patrons abstraits que nous décrivons et dont nous spécifions les processus de dérivation sont également constructibles à partir du codage des données empiriques dans une procédure ascendante. Il reste à déterminer la fréquence d'emploi de ces patrons et à analyser leur variabilité par l'étude de vastes corpus représentatifs des différents styles de parole.

5. BIBLIOGRAPHIE

Beckman, M. & Edwards, J. (1992). "Intonational categories and the articulatory control of duration", in Tokura & al., eds: *Speech Production, Perception and Linguistic Structure*, IOS Press.

Clements, G.N. & Ford, K. (1979. Kikuyu tone shifts and its synchronic consequences, *Linguistic Inquiry*, 10, 179-210.

Di Cristo, A. (1978). *De la Microprosodie à l'Intonosyntaxe*, Thèse d'Etat, Un. de Provence.

Di Cristo, A. & Hirst, D.J. (1992). "Rythme syllabique, rythme mélodique et représentation hiérarchique de la prosodie du français", *TIPA* 15, 9-24.

Di Cristo, A. & Hirst, D.J. (1993). Prosodic regularities in the surface structure of French questions, *Working Papers* (Lund), 41, 268-271.

Goodman, J.C., Lee, L & de Groot, J. (1994). "Developing Theories of Speech Perception", in Goodman & Nusbaum, eds: *The Development of Speech Perception*, MIT Press.

Hirst,D.J.(1980). "Un modèle de production de l'intonation",*TIPA* 7, 297-315.

Hirst, D.J. & Di Cristo, A. (1984). "French Intonation: a parametric approach", *Die Neueren Sprachen*, 83, 554-569.

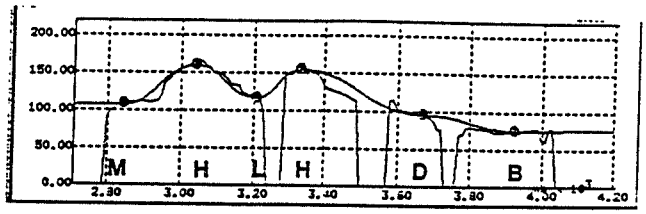
Hirst, D.J. & Espesser, R. (1993). Automatic modelling of fundamental frequency using a quadratic spline function, *TIPA*, 15, 71-85.

Nespor, M. & Vogel, I. (1986). *Prosodic Phonology*, Foris.

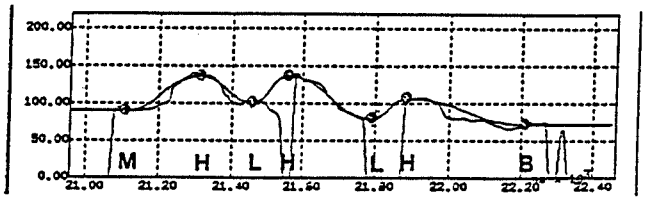
Pierrehumbert, J. (1980) *The Phonology and phonetcs of English Intonation*, Ph.D. MIT.

Selkirk, E.O. (1978/1981). On prosodic structure and its relation to syntactic structure, *Nordic Prosody*, II, 111-140.

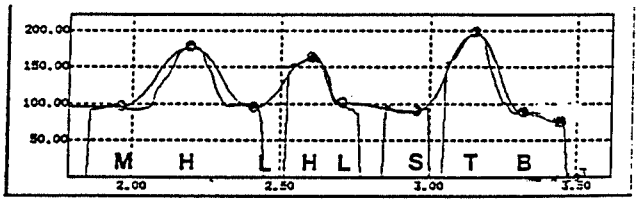
Vaissière, J. (1995). Phonetic explanations for cross-linguistic prosodic similarities, *Phonetica*, 52, 123-130.



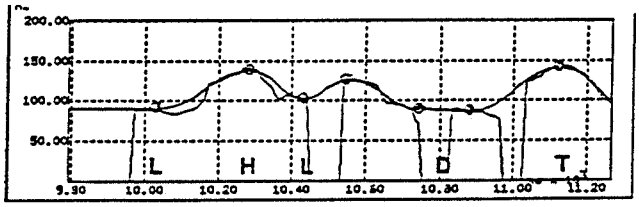
1a. Une bonne bouteille de champagne (assertif)



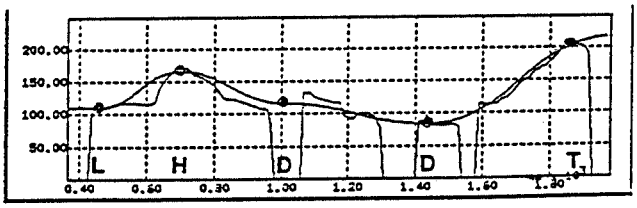
1b. Une bonne bouteille de champagne (assertif marque)



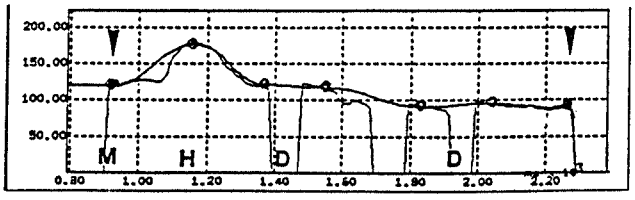
1c. Une bonne bouteille de champagne (emphatique)



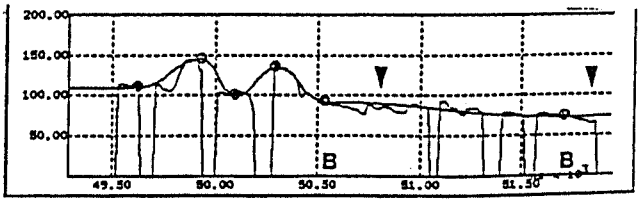
1d. Une bonne bouteille de champagne (continuatif)



1e. Une bonne bouteille de champagne (question)



1f. Une bonne bouteille de champagne (extrapose gauche)



1g. Une bonne bouteille de champagne (extrapose droite)

Figure 1 : Courbes de F0, modélisations et codage INTSINT des patrons mélodiques décrits dans le texte.

"Y A-T-IL DES UNITÉS TONALES EN FRANÇAIS?"

Daniel HIRST & Albert DI CRISTO

Institut de Phonétique d'Aix - URA CNRS 261 .Parole et Langage,
Université de Provence - 29 avenue Schuman - 13621 Aix-en-Provence
Tél.: +33 42 953628 - Fax: +33 42 595096 - email hirst@univ-aix.fr

ABSTRACT

In this paper we examine a number of arguments which have been put forward recently arguing against the inclusion of the Tonal Unit as a distinct level in the prosodic hierarchy. We show that the data which has been presented as problematic can in fact be easily accounted for in our model and we present further data which we cannot account for without postulating the existence of Tonal Units.

1. INTRODUCTION

Nous avons développé dans le passé (Hirst 1983, Hirst & Di Cristo 1984) une approche paramétrique de l'intonation de l'anglais et du français dans laquelle une représentation phonologique est formée de deux unités hierarchisées : l'Unité Tonale (**UT**) et l'Unité Intonative (**UI**) . Dans un travail plus récent (Di Cristo & Hirst 1993), nous avons exposé des arguments en faveur d'une unité phonologique de rang intermédiaire que nous avons appelée Unité Rythmique (**UR**). A titre d'exemple, la proposition initiale de l'énoncé: "Ils discutaient de sa décision quand nous sommes entrés" serait donc analysée :

(1)[<(ils dis)(cutaient)><(de sa dé) (cision)>]

où [...] représente l'**UI**, <...>, l'**UR** et (...), l'**UT**.

Dans une étude récente, qui s'inspire du cadre théorique élaboré par Pierrehumbert & Beckman (1988), Jun & Fougeron (1995) optent pour une autre interprétation. Selon ces auteurs, en effet, il n'y aurait que deux unités prosodiques en français : le "Groupe Intonatif" (intonation group) correspondant à notre **UI**, et le "Syntagme Accentuel" (accentual phrase) qu'elles identifient à notre **UR**, ce qui équivaut en fait à reprendre l'analyse en **GA** (Groupe Accentuel) et **GI** (Groupe Intonatif) adoptée antérieurement par d'autres chercheurs (Di Cristo, 1978; Verluyten, 1983, Mertens, 1993). Ainsi dans une séquence comme "le garçon malade" (leur exemple) que nous analyserions comme :

(2) <(Le gar-)(-çon malade)>

<(L H)(L H)>

Jun et Fougeron interprètent la suite LH observée sur la première **UT** comme la première partie d'un schéma global <LHLH> qui serait assigné directement au niveau du syntagme accentuel.

Nous nous proposons, dans la première partie de cette communication, d'exposer les arguments invoqués par J & F à l'encontre de notre analyse et de donner les raisons pour lesquelles ils ne nous paraissent pas convaincants. Nous présenterons ensuite nos propres arguments et des données empiriques en faveur du maintien de l'**UT** dans la représentation hiérarchique des unités prosodiques.

2. PRESENTATION ET DISCUSSION DES ARGUMENTS.

2.1. Arguments contre l'Unité Tonale

Jun et Fougeron soutiennent l'idée qu'il existe un seul constituant prosodique de rang inférieur à l'**UI**: le Syntagme Accentuel dont le début et la fin sont marqués tonalement. Cette unité est en fait très proche de "l'arc accentuel" de Fónagy (1980) ou du mot phonologique de Milner & Regnault (1984).

Les deux principaux arguments de Jun et Fougeron à l'encontre du statut phonologique de l'UT sont les suivants :

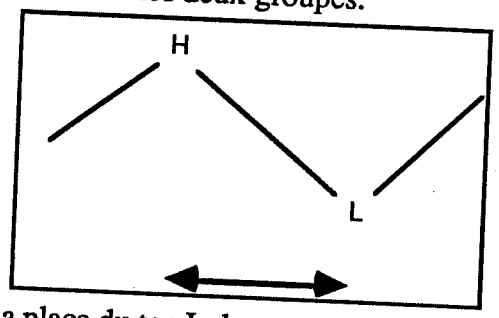
(i) Si on convient que le mot fait partie des catégories phonologiques, un modèle incluant des Unités Tonales ne respecterait pas l'hypothèse de la hiérarchie stricte ("Strict Layer Hypothesis, Selkirk 1978), selon laquelle chaque niveau est constitué d'une séquence d'éléménts de niveau *strictement* inférieur. En effet, s'il est toujours possible de décomposer un syntagme accentuel (ou **UR**) en une suite de mots :

(3). <le+garçon+malade>

il n'en va pas de même pour l'**UT** dont les frontières ne tiennent pas compte des limites des mots.

(ii) La durée de la chute comprise entre le premier pic (ton H) du syntagme accentuel et le creux suivant (ton L) est proportionnelle au

nombre de syllabes dans le syntagme. Dans le cas de deux groupes accentuels, par contre, ce 'temps de chute' entre le dernier pic du premier groupe accentuel et le premier creux du second serait constant, quel que soit le nombre de syllabes dans les deux groupes.



La place du ton L dans un schéma comme le précédent serait donc dépendent de la nature de la frontière prosodique située entre les deux tons.

2.2. Discussion des arguments.

La critique selon laquelle notre analyse ne respecte pas "l'hypothèse des niveaux stricts" ne nous paraît pas décisive. En effet, on sait depuis longtemps que cette hypothèse impose des contraintes trop exigeantes sur les représentations phonologiques. Ladd (1986), par exemple, avance des arguments convaincants pour l'emploi de catégories récursives en phonologie comme en syntaxe, où une catégorie X domine un autre élément qui lui-même est de catégorie X. De telles structures sont évidemment en contradiction flagrante avec l'hypothèse des niveaux stricts.

On sait également que le statut phonologique du 'mot' est l'objet de controverses. En effet, si on accepte la syllabe comme unité phonologique, il n'est pas possible d'établir une hiérarchie stricte qui contiennne à la fois la syllabe et le mot. Ainsi, dans la séquence "il est en or", le découpage syllabique /i.le.ta~.nø‰/ ne respecte pas un découpage morpho-syntaxique en mots. Du reste il est établi que la formation des pieds métriques ne respecte pas non plus les frontières de mots (Abercrombie 1964).

On concluera donc sur ce point que si l'hypothèse des niveaux stricts soulève des problèmes pour l'**UT**, elle en pose également pour la syllabe et pour le groupe accentuel (ou pied métrique) qui sont pourtant très généralement acceptés comme des constituants prosodiques universels.

(ii) A la différence du premier argument théorique qui vient d'être discuté, l'expérience rapportée par Jun et Fougeron est fondée sur des données empiriques intéressantes dont tout modèle adéquat doit pouvoir rendre compte. Comme on va le voir, ces résultats ne posent pas de réels problèmes pour notre modèle.

Si nous acceptons l'équivalence proposée par Jun et Fougeron entre le Syntagme Accentuel et l'**UR**, nous pouvons résumer leurs observations de la façon suivante : le temps de chute entre un ton H et le ton L subséquent n'est pas le même dans les exemples (4a (et (4b):

(4) a. (... H)(L...)

b. <...(... H)><(L ...)...>

Plus spécifiquement, Jun et Fougeron constatent que dans le cas (4a) (une **UR** contenant deux **UT**) la position du ton L après la frontière est dépendante du nombre de syllabes de la deuxième **UT** : ce ton L semble donc être aligné (comme le deuxième ton H) avec la **fin** de la deuxième unité. Dans le cas (4b) par contre, (deux **UR** contenant chacune une **UT**), on n'observe pas de corrélation entre la position du deuxième ton L et le nombre de syllabes de la deuxième unité. Le deuxième ton L semble donc être aligné, contrairement au deuxième ton H, avec le **début** de la deuxième unité.

Ces résultats, malgré leur intérêt évident, ne constituent pas un argument contre l'existence des **UT**. Nous pourrions, en effet, intégrer des principes d'interprétation phonétique pour rendre compte de façon appopriée de l'alignement des tons dans les structures (4a) et (4b). Il s'avère toutefois que de telles règles d'interpétation ne sont pas nécessaires. Si nous regardons de plus près les exemples donnés par Jun et Fougeron à titre d'illustration : "Marmonnement..." et "Le désagréable garçon..." (extraits des phrases "Marmonnement est un mot utilisé par les français" et "Le désagréable garçon ment à sa mère."), nous constatons que nous ne pouvons pas accepter l'équivalence proposée entre leur Syntagme Accentuel et notre **UR**. Dans les deux cas, nous analyserions ces séquences comme constituant des **UI** à part entière. En effet, on voit clairement sur les figures qui illustrent leur travail que la montée finale atteint un niveau nettement supérieur à celui de la montée initiale. L'émergence de cette montée finale constitue la marque diagnostique d'une frontière d'**UI**, l'abaissement successif des montées étant un indice incontournable de la cohésion tonale observée à l'intérieur d'une **UI**. On remarquera en outre que dans les deux exemples proposés, il s'agit d'un syntagme nominal sujet qui est assez généralement en français le lieu d'une frontière d'**UI** (Di Cristo 1978, Hirst & Di Cristo 1984).

Si nous regardons les prédictions de notre modèle pour les structures de ce type, nous trouvons en fait une explication plausible des résultats de Jun et Fougeron. La phrase "Le désagréable garçon ment à sa mère" aura la structure :

(5) [<(le dé-)(-sagréable)><(garçon)>]
[<(ment)><(à sa mère)>]

En appliquant les règles d'affectation tonale appropriées (cf Di Cristo & Hirst 1996 pour un résumé), on obtient les représentations tonales soujacentes suivantes :

(6) [L<(L H)(L H)><(L H)>H]
[L<(L H)><(L H)>L]

L'application des principes de linéarisation de simplification, d'abaissement et d'interprétation permet de convertir cette séquence en une représentation phonologique de surface de type INTSINT (Hirst & Di Cristo à paraître, Di Cristo & Hirst 1996) :

(7) [<(M H)(D)><(D T)>]
[<(M H)(D B)>]

(où M= Mid, D= Down, T = Top).

Cette représentation est parfaitement compatible avec l'exemple donné par Jun et Fougeron.

On constate que la durée mesurée par ces auteurs, dans le cas où elles analysent la séquence: "Le désagréable garçon ment à sa mère" comme deux syntagmes accentuels, est celle entre le ton (T) de la frontière finale ("garçon") de la première **UI** et le ton (M) de la frontière initiale ("ment") de la deuxième **UI**. Il n'est donc pas surprenant que ce dernier soit localisé par rapport à la frontière initiale de l'unité.

3. ARGUMENTS EN FAVEUR DE L'UNITÉ TONALE.

Nous pensons avoir montré que la ré-analyse proposée par Jun et Fougeron ne rend pas mieux compte des données que ne le fait notre analyse originelle. Il nous reste, d'une part, à montrer que celle-ci donne lieu à des généralisations qui ne sont pas effectuables dans le cadre proposé par ces auteurs et, d'autre part, à valider empiriquement nos propositions

Une analyse qui n'utiliserait pas l'**UT** comme constituant soulève deux problèmes essentiels

Premièrement, le fait d'analyser le syntagme comme une seule unité marquée à son début et à sa fin implique qu'il ne peut y avoir plus de deux mouvements mélodiques à l'intérieur du syntagme accentuel. Selon notre analyse, au contraire, rien n'interdit la formation d'une **UR** contenant trois **UT**, comme, par exemple, dans une lecture possible de l'expression "la poly-syllabicité" avec deux accents initiaux de morphèmes, que nous représenterions :

(8) <(la po-)(-ly-syll-)(-abicité)>

Une telle séquence serait impossible à représenter sans un remaniement important dans le cadre proposé par Jun et Fougeron. Ces auteurs seraient en effet obligés d'introduire un contour spécial <LHLHLH> pour rendre compte de ce cas. Seule l'étude de corpus étendus pourra dire si, comme nous le pensons, de telles séquences sont effectivement observées de façon assez systématique dans la parole spontanée.

Un deuxième inconvénient de cette analyse provient selon nous du fait qu'il n'existe plus dans la représentation de marque formelle indiquant la place du début du mot lexical. Or, selon les propos de Jun et Fougeron elles-mêmes, l'accent initial survient quelquepart entre la première et la troisième syllabe du mot lexical.

Nous avons réalisé une expérience pour vérifier si la localisation du premier pic de F0 était ou non liée au début du mot lexical ou si au contraire, elle était alignée avec le début du syntagme accentuel. Pour ce faire, nous avons pris les deux mots "directeur" et "manifestation" que nous avons fait précéder d'une, deux ou trois syllabes dans les expressions :

(a) Au directeur ...Une manifestation ...

(b) A ces directeurs...Par les manifestations...

(c) Il était directeur ...C'était une manifestation...

Ces expressions ont été insérées dans des phrases qui ont été produites par 7 locuteurs natifs français comme des réponses à des questions. Nous n'avons donné aucune consigne concernant l'accentuation des mots. Les phrases tests ont été mélangées à d'autres phrases utilisées comme distracteurs. Nous avons éliminé 4 cas où un locuteur n'a pas réalisé d'accent initial sur le syntagme. Nous avons également éliminé 5 phrases où un accent supplémentaire avait été introduit sur le mot "était". Pour les autres phrases, les enregistrements ont été numérisés et les courbes de fréquence fondamentales ont été modélisées avec le système MOMEL (Hirst & Espesser 1993). Nous avons relevé pour chaque phrase la distance entre le début du voisement et le premier point cible représentant un maximum local en début de phrase.

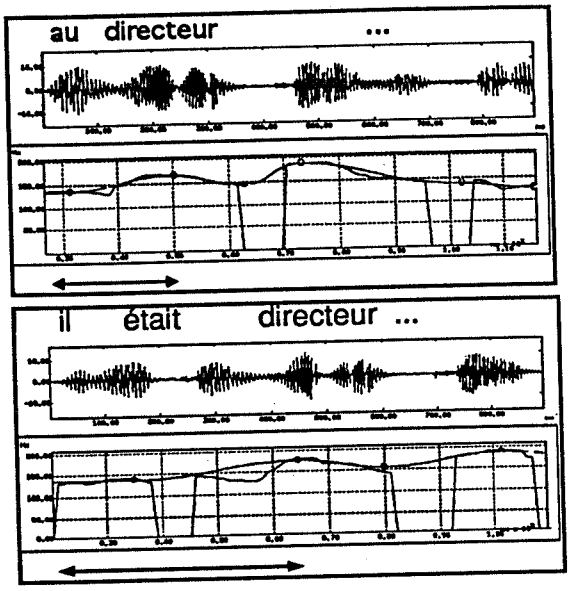


Figure 1 : Courbes modélisées (points cibles) illustrant deux exemples de notre corpus. Les doubles flèches indiquent la distance mesurée entre le début du voisement et le premier pic de F0.

Une analyse de variance montre un effet très significatif du nombre de syllabes inaccentuées sur la position du premier pic ($F(2,31) = 5,545$; $p = 0,0087$). Nous avons effectué, d'autre part, une régression linéaire qui nous fournit l'équation :

$$Dist = 111 * nsyll + 259$$

avec un taux de significativité encore plus élevé ($p = 0,0021$).

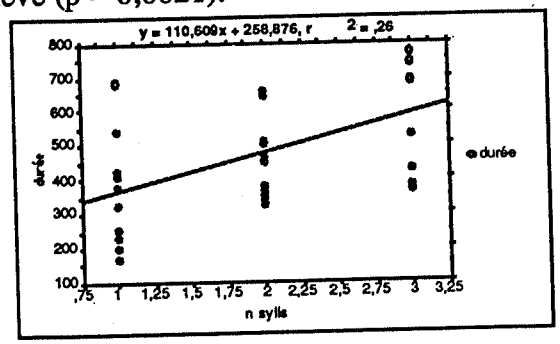


Figure 2 : régression linéaire entre le nombre de syllabes inaccentuées et la position du pic initial par rapport au début de voisement.

Au vu du résultat de cette régression, nous pouvons conclure que le premier pic est bien localisé en fonction de la place de la première syllabe du mot lexical.

4. CONCLUSION

Nous avons présenté des éléments de discussion concernant la validité du constituant Unité Tonale dans un modèle phonologique de l'intonation du français. Cette discussion a donné lieu à une confrontation des prédictions de notre approche avec celles récemment exposées par Jun & Fougeron (1995), qui tentent d'appliquer au français les principes d'analyse de Beckman & Pierrehumbert (1986). Nous pensons avoir montré qu'un modèle qui ne prend pas en compte l'Unité Tonale limite la portée des généralisations que l'on peut effectuer sur la prosodie du français. D'autre part, les résultats des expériences pilotes que nous avons effectuées sont de nature à confirmer la nécessité d'intégrer l'UT dans l'organisation intonative de cette langue.

BIBLIOGRAPHIE

Abercombrie, D. (1964). "Syllable quantity and enclitics." in Abercrombie,et al. (eds) *In Honour of Daniel Jones*. Longmans, Londres. 216-222.

Beckman, M. & Pierrehumbert, J. (1986). "Intonational structure in Japanese and English" *Phonology Yearbook*, 3, 255-309.

Di Cristo, A. (1978).*De la Microprosodie à l'Intonosyntaxe*, Thèse d'Etat. Université de Provence.

Di Cristo, A. & Hirst, D.J. (1993). "Rythme syllabique, rythme mélodique et représentation hiérarchique de la prosodie du français", *TIPA*, 15, 9-24.

Di Cristo, A. & Hirst, D.J. (1996) "Vers une typologie des Unités Intonatives en français." Communication à ce Colloque.

Fónagy, I. (1980). "L'accent en français: accent probabilitaire". *Studia Phonetica*, 15, 123-233.

Hirst, D.J. (1983). Structures and categories in prosodic representations, in Cutler, A. & Ladd, D.R. *Prosody: Models and Measurements*, 93-109.

Hirst, D.J. & Di Cristo, A. (1984). "French intonation: a parametric approach". *Die Neueren Sprachen*, 83, 554-569.

Hirst, D.J.; Di Cristo, A. (à paraître), "A survey of intonation systems", in *Intonation systems : a survey of twenty languages*, Cambridge University Press.

Hirst, D.J. & Espesser, R. (1993). "Automatic modelling of fundamental frequency using a quadratic spline function" *TIPA*, 15, 71-85.

Jun, Sun-Ah & Fougeron, C. (1995). "The accentual phrase and the prosodic structure of French", *Proc. ICPhS* Stockholm, 2, 722-725.

Ladd, D.R. (1986). "Intonational phrasing. The case for recursive structures". *Phonology Yearbook*, 3, 311-340.

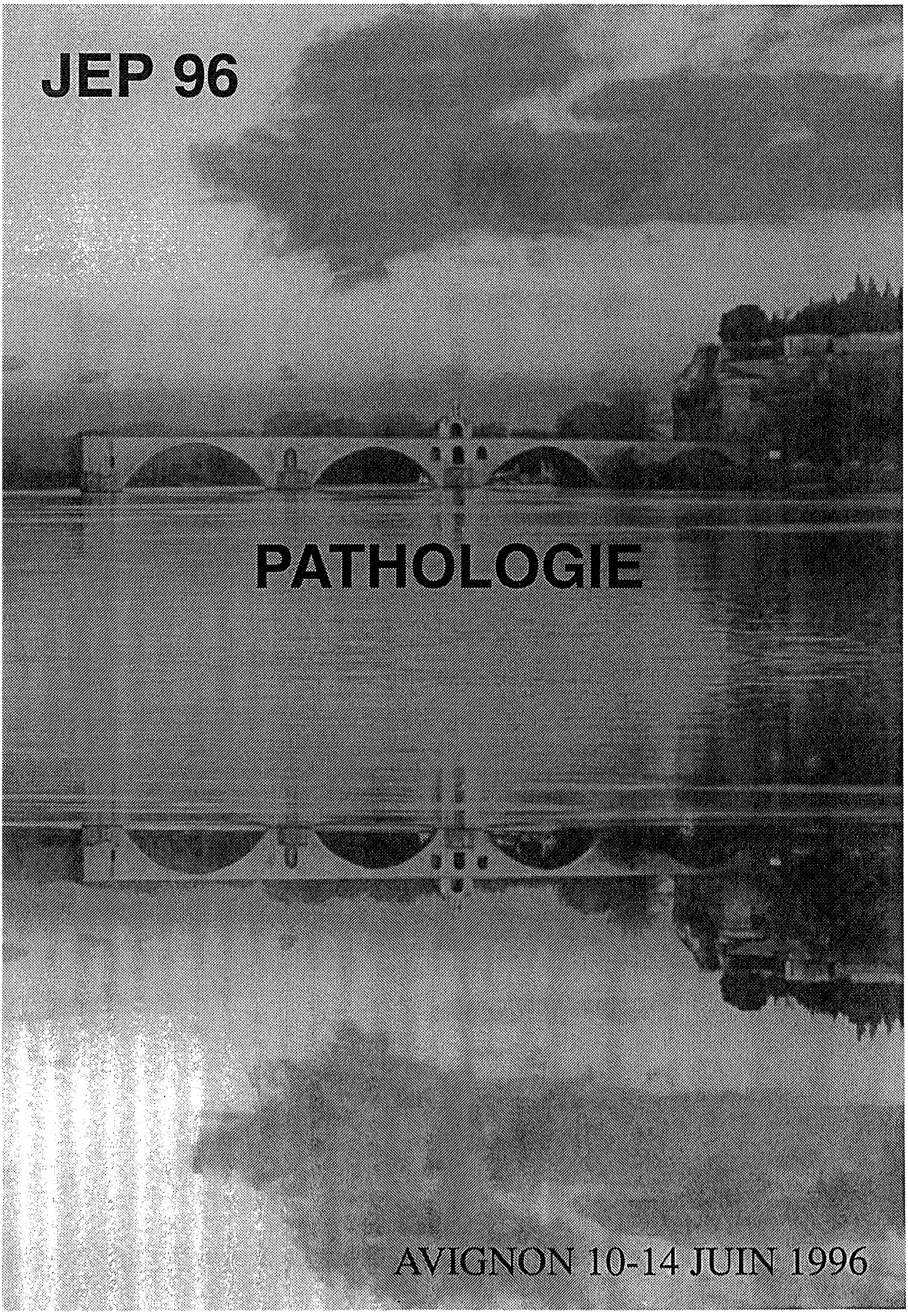
Mertens, P. (1993). Accentuation, intonation et morphosyntaxe, *Travaux de Linguistique*, 26, 21-69.

Milner, J.C. & Regnault, F. (1987). *Dire le Vers*. Seuil.

Selkirk, E.O. (1978/1981). On prosodic structure and its relation to syntactic structure, in: Fretheim (ed). *Nordic Prosody* II, 111-140.

Verluyten, P. (1983). La structuration de l'énoncé en groupes intonationnels, *ITL Review of Applied Linguistics*, 60-61, 77-104.

JEP 96



PATHOLOGIE

AVIGNON 10-14 JUIN 1996

ÉVALUATION OBJECTIVE DES RÉGLAGES D'UN IMPLANT COCHLÉAIRE PAR ANALYSES DISCRIMINANTES STATISTIQUES.

Yvan LIMOUSI, Willy SERNICLAES

Laboratoire Audioprothèse Implantée, Hopital StAntoine,184 rue du Fbg St-Antoine 75 012 Paris
Tél : 49 28 20 00, poste 4400. Fax : 49 28 24 95
Institut de phonétique, 19 rue des Bernardins 75 005 Paris
URA CNRS *Landisco*, Sciences du Langage, U. Nancy II
Ecole de Santé Publique, Univ. Libre de Bruxelles

1. INTRODUCTION :

Les implants cochléaires sont des prothèses auditives destinées aux sourds profonds. Contrairement aux prothèses classiques, leur mise en place nécessite une intervention chirurgicale au cours de laquelle 15 électrodes sont placées à l'intérieur de la cochlée. Ces électrodes ont pour fonction de transmettre au nerf auditif l'information acoustique transformée par un processeur en intensités électriques. Chaque électrode couvre une région du spectre dont la largeur de bande est définie selon une "stratégie" de réglage.

La méthode d'évaluation des réglages d'un implant cochléaire par analyses statistiques a pour objectif essentiel d'apporter un gain de temps dans les manipulations effectuées par le praticien ORL après la mise en place de l'implant. Elle doit également permettre de définir le meilleur ajustement possible. En effet, la technologie offre maintenant pour la plupart des implants cochléaires la possibilité d'intervenir de manière post- opératoire sur le réglage du processeur qui effectue le traitement du signal acoustique. Ce progrès offre un avantage considérable car il permet d'adapter les réglages en fonction du profil de chacun des patients et de les modifier jusqu'à obtenir un résultat satisfaisant pour le patient implanté. Cependant, la mise en oeuvre du processus de réglage peut être longue et hasardeuse pour deux raisons :

En premier lieu, le praticien ne dispose que de l'appréciation du patient pour déterminer quelles caractéristiques du réglage doivent être modifiées. Si pour certains aspects, tels que la détermination des seuils de confort, l'appréciation du patient peut être considérée comme nécessaire et suffisante, il n'en va pas de même pour d'autres réglages plus subtils qui interviennent sur la qualité de la transmission des données. Par exemple, le module de réglage de l'implant cochléaire DIGISONIC contient une partie essentielle permettant de définir la répartition du spectre fréquentiel à travers les 15 électrodes. Afin d'apprécier l'effet obtenu par une modification de cette répartition sur la qualité de la transmission, il est nécessaire de comparer entre une série de choix possibles. Cependant, des tests de perception montrent que les performances auditives d'un patient implanté augmentent en fonction de l'adaptation du patient à une certaine stratégie de réglage. L'effet d'habituation à une stratégie réglage soulève dès lors des problèmes méthodologiques car le patient aura tendance à sous-estimer les apports d'une nouvelle stratégie de réglage.

La deuxième raison à l'aspect hasardeux de la démarche de réglage est qu'il est difficile de déterminer quelles modalités du réglage doivent être modifiés pour obtenir une réelle amélioration de la transmission. Il est plus prudent, lorsque cela est possible de s'appuyer sur des données objectives fondées sur des réalités acoustiques et qui permettent de prendre des décisions pertinentes.

Ces inconvénients limitent donc les possibilités d'améliorer le réglage d'un implant cochléaire. C'est pourquoi l'utilisation d'une méthode objective permettant de déterminer un réglage optimal par classification automatique de sons de parole trairtés par l'implant s'avère très avantageuse.

La méthode objective proposée par W. Serniclaes consiste en une analyse discriminante statistique. Des recherches antérieures (W. Serniclaes) ont déjà permis de comparer deux stratégies de réglage (stratégie linéaire : à chacune des 15 électrodes, une bande fréquentielle de largeur égale est attribuée; stratégie Mel : les seuils des bandes de fréquences sont équidistants sur l'échelle des Mel). Les observations faites sur les résultats obtenus avec ces stratégies ont permis de détecter quelles régions du spectre semblaient être les plus informatives. Cette information a donné naissance à une nouvelle stratégie. Celle-ci offre un examen plus approfondi de la région des basses et moyennes fréquences. De plus, elle simule la correction de la perte de 6 dB par octave jusqu'à 4 kHz opérée par l'oreille. Cette

nouvelle stratégie (Mel Modified HS) a été testée avec succès par la méthode d'évaluation objective sur un corpus de consonnes et de voyelles (Y. Limousi), (S. Gérardot). Cependant, la pertinence dans le choix de cette méthode se devait d'être démontrée. En effet, le but de la méthode objective est de simuler la reconnaissance perceptive. Par conséquent, les résultats qu'elle permet d'obtenir doivent être comparables à ceux de la perception auditive.

L'objet de cette recherche est de tester la pertinence perceptive de la méthode d'évaluation objective des implants par analyses statistiques. À cette fin, les résultats de classification d'un corpus de consonnes seront comparés à ceux qui ont été observés dans les tests de perception des consonnes par Miller et Nicely (Miller et Nicely 1955).

2. LA MÉTHODE OBJECTIVE :

La méthode objective consiste à utiliser une analyse discriminante statistique pour classer des segments de parole en catégories phonémiques. Les phonèmes classés dans cette recherche sont 16 consonnes du français (p, t, k, b, d, g, f, s, ch, v, z, j, m, n, l, r). Le corpus utilisé est constitué de 61 mots monosyllabiques CV ou débutant par l'une des 16 consonnes suivie d'une des 4 voyelles « a, i, ou, an ».Les 61 mots ont été prononcés à trois reprises par 10 locuteurs (61*3*10=1830 stimuli). Chacun des mots a été segmenté manuellement en 4 parties. Ces parties reflètent 4 étapes articulatoires de la séquence acoustique, à savoir :

- la partie antérieure à la détente des occlusives ou la constriction des fricatives.
- la détente d'occlusion ou la fin de constriction.
- les transitions
- la partie vocalique stable.

Les sonantes « l, m et n », ont été segmentées sur le modèle des occlusives. « r » a été segmentée comme une fricative.

Les moyennes des intensités pour chacun des 4 segments et des 15 bandes de fréquences correspondant au 15 électrodes de l'implant (4*15 = 60 données par mot) ont été utilisées comme indices pour classer les consonnes en 16 groupes à l'aide de l'analyse discriminante.

Afin de comparer nos résultats avec ceux des tests de perception de Miller et Nicely, nous avons procédé à une classification en fonction des 4 traits phonétiques (lieu d'articulation, mode d'articulation, voisement et nasalité). Le trait de durée n'a pas été examiné car seules les moyennes d'intensité et non pas la durée des segments ont été utilisées dans l'analyse discriminante. Une simulation de filtrage passe-haut et passe-bas a été effectuée en supprimant successivement les indices fournis soit par les fréquences élevées, soit par les fréquences basses

Tableau 1 : Attribution des traits phonétiques aux consonnes du corpus :

	p	t	k	b	d	g
lieu d'art.	0	1	2	0	1	2
mode d'art	0	0	0	0	0	0
nasalité	0	0	0	0	0	0
voisement	0	0	0	1	1	1

	f	s	ch	v	z	j
lieu d'art.	0	1	2	0	1	2
mode d'ar	1	1	1	1	1	1
nasalité	0	0	0	0	0	0
voisement	0	0	0	1	1	1

	l	m	n	r
lieu d'art	2	0	1	2
mode d'ar	2	2	2	2
nasalité	0	1	1	0
voisement	1	1	1	1

-Lieu d'articulation : labiales = 0, coronales = 1, vélaires = 2.

-Mode d'articulation : occlusives = 0, fricatives = 1, sonantes = 2.

-Voisement : voisées = 0, non voisées=1.

-Nasalité : nasales =0, non nasales = 1.

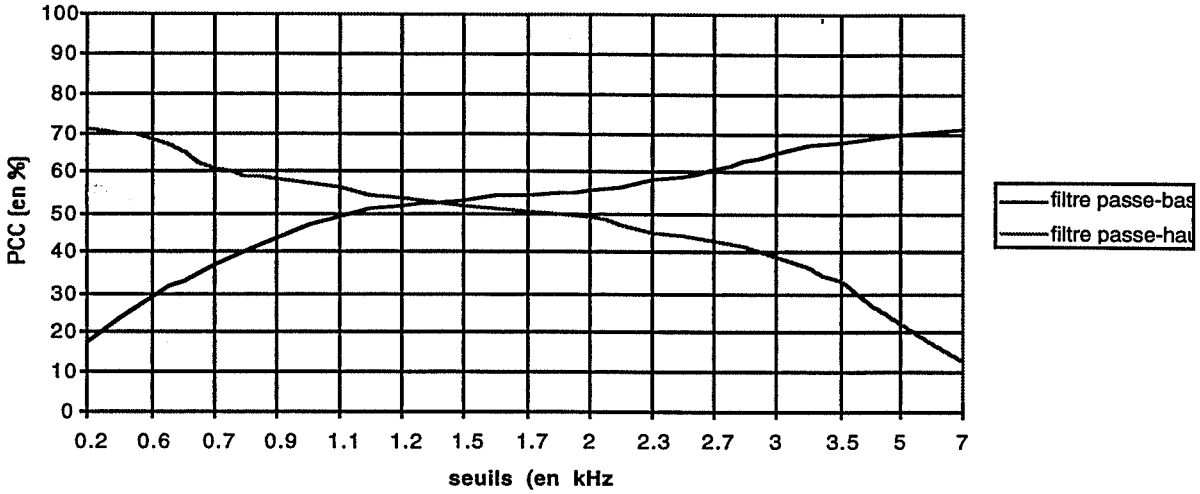
Les pourcentages de classification correcte (PCC) par l'analyse discriminante permettent d'évaluer le bien fondé d'une stratégie de réglage. C'est la nouvelle stratégie (MM HS) qui a servi de support aux observations de la présente recherche.

3. RÉSULTATS :

Tableau 2 : Pourcentage de classification correcte par l'analyse discriminante statistique

	Cons.	Lieu	Mode	Nasal.	Voise.
PCC (%)	71,0	65,30	73,99	84,15	92,62

PCC des consonnes avec filtre passe-haut et filtre passe-



4. DISCUSSION :

La première observation que l'on peut faire sur ces résultats est leur relative médiocrité en regard des résultats obtenus dans des tests de perception de Miller et Nicely (90,85 % avec S/N = 12dB et filtrage bande passante 200-6500 cps). Ceci s'explique par le fait que l'information utilisée par l'analyse discriminante statistique est bien moins importante que celle dont dispose un sujet pour effectuer une tâche de perception, car toute l'information qui est normalement transmise par les milliers de cellules ciliées doit être regroupée ici en 15 électrodes. Mais cet aspect est intrinsèquement lié aux limites technologiques et physiologiques de l'implantation cochléaire. Un autre aspect, celui-là, propre à la méthode proposée est que l'information temporelle n'est rendue que très sommairement par une segmentation minimale des stimuli.

Cependant, l'essentiel n'est pas tant de considérer la quantité de réponses correctes que l'évolution des résultats en fonction des conditions de transmission, car si cette évolution ressemble à celle observée en perception, il sera permis de dire que la méthode objective simule correctement la perception. Cette assertion permettra de considérer que toute amélioration des scores de classifications (obtenus par l'analyse discriminante) d'une stratégie de réglage par rapport à une autre sera une preuve de sa plus grande efficacité lors de son utilisation sur un implant cochléaire.

Les résultats obtenus permettent de faire trois observations :

Le premier critère permettant de montrer que la méthode objective opère de la même façon que la perception s'obtient en comparant les résultats objectifs et perceptifs dans conditions dégradée. Dans notre expérience, un filtrage passe-bas et passe-haut ont été simulés en supprimant progressivement les informations portées par les électrodes correspondant à la fréquence désirée. Le graphique 1 montre que le point d'intersection des deux courbes représentant les résultats avec les filtrages passe-bas et passe-haut pour la classification des consonnes se situe entre 1,2 et 1,5 kHz (1550 Hz en perception (W. Serniclaes)). Ce point d'intersection montre que l'ensemble des informations se situant sous cette fréquence a autant de valeur que celles qui se situent au-dessus. Il se trouve très bas puisqu'il correspond au 1/5ème de la plage totale des fréquences. Ceci montre que l'information la plus discriminante se situe dans une région limitée des basses fréquences, comme l'ont observé Miller et Nicely dans leurs tests de perception.

Le deuxième critère est l'observation des matrices de confusion. Celles-ci nous permettent de voir que non seulement les erreurs sont en conformité avec le modèle des traits phonétiques (confusions plus souvent faites entre consonnes séparées par un seul trait phonétique) mais qu'elles tendent à le demeurer lorsque les hautes fréquences sont supprimées et au contraire à ne plus l'être lorsque les basses fréquences sont occultées. Cette observation rejoint celles de Miller et Nicely qui ont montré l'importance des basses fréquences pour une bonne discrimination acoustique.

Enfin, on observe que les traits phonétiques les plus résistants et les mieux reconnus sont les

mêmes qu'en perception, à savoir dans l'ordre : voisement, nasalité, mode, lieu. Cette comparaison doit cependant être relativisée car les études de Miller et Nicely ont été effectuée sur un corpus de consonnes anglaises définies par des traits phonétiques qui n'ont pas toujours les mêmes manifestations acoustiques qu'en français (ex : pas d'aspiration comme manifestation du trait non voisé des occlusives françaises)

Ces différentes observations montrent que la méthode objective par analyses statistiques discriminantes utilise les données dont elle dispose de manière pertinente, de sorte que son comportement est très proche de celui de la perception normale. On peut donc considérer qu'elle fournira une évaluation fiable des stratégies de réglage, au moins lorsqu'il s'agira d'en comparer plusieurs entre elles. Dans l'immédiat, un implant cochléaire tel que le DIGISONIC peut utiliser la méthode objective pour définir laquelle des trois stratégies (linéaire, Mel, Mel HS) est objectivement la meilleure pour un patient en fonction de ses seuils. Mais cette méthode est également un outil qui permet d'aider à élaborer de nouvelles stratégies encore mieux adaptées au patient. Ceci est fortement suggéré par l'amélioration des résultats obtenus avec la nouvelle stratégie MM HS.

BIBLIOGRAPHIE

W. Serniclaes et al : « *Objective Evaluation of Vowel Identification with the Digisonic Cochlear Implant* », Audiology, à paraître.

G.A. Miller & P.E. Nicely (1955) : « *An analysis of Perceptual Confusions Among some English Consonants* », J. Acoust. Soc. Am..

Y. Limousi : « *Evaluation objective des performances de l'implant cochléaire DIGISONIC par analyse discriminante statistique sur un corpus de consonnes* » Mémoire de Maîtrise, Nancy2.

S. Gérardot : «*Evaluation objective des performances de l'implant cochléaire DIGISONIC par analyse discriminante statistique sur un corpus de voyelles* » Mémoire de Maîtrise, Nancy2.

EVALUATION SUBJECTIVE DE LA VOIX ET DE LA PAROLE APRES LARYNGECTOMIE PARTIELLE SUPRA-CRICOIDIENNE

Lise Crevier-Buchman, Jacqueline Vaissière°, Ollivier Laccourreye Daniel Brasnu

Laboratoire de recherche sur la Voix, les Biomatériaux et la Cancérologie ORL, Hôpital Laënnec, Université Paris V, 42 rue de Sèvres, 75007 Paris. Tel: (1) 44 39 66 57, Fax: (1) 44 39 66 46.
°Institut de Phonétique de Paris, Université Paris III, UA1027, 19 rue des Bernardins, 75006 Paris. Tel: (1) 43 26 26 07, Fax: (1) 43 29 70 13.

ABSTRACT

A prospective perceptual evaluation of voice and speech after supra-cricoid partial laryngectomy over the first post-operative six months was performed. Ten male patients were evaluated at one, three and six months by five listeners. A voice profil was drawn from 12 defined acoustical parameters studied, and a chronology in the voice improvement was observed.

INTRODUCTION

De nombreux travaux ont précisé la place et l'apport de la laryngectomie partielle supracricoïdienne (LPSC) avec crico-hyoïdo-épiglottopexie (CHEP) ou crico-hyoïdopexie (CHP), dans les cancers épidermoïdes à point de départ glottique (Piquet, 1974 & 1986; Laccourreye, 1990). Le principe des LPSC était de conserver la continuité des voies aériennes malgré une exérèse large du larynx. Le but est obtenu par une impaction du cartilage cricoïde sur le cartilage hyoïde (Piquet, 1974 & 1986; Laccourreye, 1990). Par ailleurs la restauration vocale est rendue possible grâce à la conservation d'au moins une unité crico-aryténoïdienne et son innervation récurrentielle (Guerrier, 1987).

Les conséquences de la LPSC sur l'appareil phonatoire sont, d'une part une néo-glotte de masse et de volume augmenté (Crevier-Buchman, 1995; Pech, 1988 & 1990), cette néo-glotte étant constituée, en arrière par le ou les aryténoïdes, et en avant, soit par la base de langue dans la CHP (Piquet, 1986), soit par un morceau d'épiglotte dans les CHEP (Piquet, 1974). De plus, il existe un raccourcissement du tractus vocal et une modification de la forme et de la taille des cavités pharyngées (Guerrier, 1987; Crevier-Buchman, 1995; Pech, 1990). Ces modifications sont responsables d' une altération des paramètres de la voix qui est aggravée et irrégulière (Crevier-Buchman, 1995).

On retrouve dans la littérature, des évaluations objectives et subjectives de ces voix, dans le but d'en apprécier les caractéristiques acoustiques générales à un instant donné (Crevier-Buchman, 1995; Pech, 1988 & 1990; Decroix, 1976; Traissac, 1992).

Nous avons cherché à préciser l'évolution temporelle de la récupération vocale par une étude prospective sur 6 mois, de dix patients opérés d'une LPSC. Tous les patients ont été évalués par jury d'écoute dans le but d'établir un profil vocal et de préciser la chronologie de la réapparition des différents composants pertinants vocaux.

MATERIEL ET METHODE

Dans notre étude prospective, dix patients de sexe masculin, agés de 51 ans à 74 ans (médiane: 65 ans), ont tous été opérés d'une laryngectomie partielle supracricoïdienne avec 4 reconstructions par crico-hyoïdopexie (CHP) et 6 crico-hyoïdo-épiglottopexies (CHEP). Ils ont été enregistré à 1, 3, et 6 mois post-opératoires.

Les enregistrements étaient réalisés en pièce non insonorisée, sur magnétophone DAT Sony 60ES (SoftADS, SonyFrance), avec un microphone a électret Lem EMU 4535 (Lem Communication, Igny, France). Les sujets se tenaient à 15 cm du microphone.

Le protocole d'enregistrement comprenait la lecture d'un texte standard, la lecture d'une liste de mots, 5 voyelles tenues prolongées et de la parole spontanée.

Les enregistrements étaient ensuite présentés de façon anonyme au jury d'écoute. Notre jury était composé de 5 auditeurs expérimentés. A partir d'un matériel sonore enregistré au préalable, chaque auditeur remplissait une grille d'évaluation. Cette grille d'évaluation comprenait les cinq paramètres de l'échelle GRBAS (Grade, Rough, Breathy, Asthenic, Strained), pour l'analyse du timbre (rauque et soufflé) et pour l'analyse du comportement phonatoire (hypotonique et hypertonique). En complément, nous avons rajouté 7 paramètres: l'intensité en voix conversationnelle, la hauteur, la mélodie, l'instabilité du voisement, le débit phonatoire et l'intelligibilité de la parole; ces paramètres nous ont semblé incontournables dans le cadre d'une évaluation de la voix après laryngectomie partielle. Enfin, nous avons aussi rajouté une analyse comparative en lecture et en parole spontanée pour évaluer l'utilisation de la parole en situation spontanée et contrainte. Pour chaque paramètre 4 choix de cotation étaient proposés: de 0 (normal) à 3 (très altéré).

Pour contrôler la fiabilité du jury d'écoute, nous avons représenté une deuxième fois un enregistrement à l'insu des auditeurs et vérifier la concordance des résultats.

RESULTATS

Les résultats de l'évaluation subjective par jury d'écoute, de la voix et de la parole des patients opérés d'une LPSC reconstructive, au cours des trois enregistrements post-opératoires sont présentés Tableau I. On observe une amélioration de 8 paramètres sur 12, toutefois, certains changements sont plus notables que d'autres (marqués par +++ sur le tableau).

A partir des résultats des grilles d'évaluations, nous avons établi un profil vocal des sujets à 1, 3 et 6 mois de leur opération. Ce profil vocal concerne l'absence d'altération (cotation 0) de chacun des paramètres étudiés. Le Tableau I représente le « profil vocal positif » des patients.

DISCUSSION

De nombreux auteurs (Crevier-Buchman, 1994; Pech, 1988 & 1990; Decroix, 1976; Traissac 1992; Arnoux-Sindt, 1992) ont signalés l'amélioration progressive des paramètres acoustiques de la voix et de la parole après laryngectomie partielle supra-cricoïdienne. Nous avons cherché a préciser quelle était l'évolution chronologique de la récupération des différents paramètres acoustiques de la phonation, au travers d'une analyse par jury d'écoute.

Pour l'analyse subjective en pathologie, se pose le problème de la terminologie employée par les spécialistes de la voix. En effet, on rencontre souvent des divergences dans les qualifications des troubles, à cause d'une référence acoustique qui se révèle très personnelle (Bassich, 1986) .

Ainsi, plusieurs essais pour standardiser les tests d'évaluation de la voix humaine ont été réalisés, dans le but d'utiliser une terminologie qui renvoie à des définitions précises pour l'ensemble des équipes prenant en charge les problèmes de voix (Fex, 1992). Hirano (1989) a proposé un modèle d'évaluation du timbre et du comportement vocale: le GRBAS. Le GRBAS semble être un des protocoles standardisés les plus répandus. La fiabilité et la reproductibilité de la méthode ont été éprouvées pour des voix pathologiques en général (Dejonckère, 1993). Nous avons donc eu recours a cette échelle d'évaluation en y ajoutant certains paramètres qui nous ont semblés pertinent dans la description des voix apres LPSC. Il s'agissait de la qualité du voisement, de la hauteur tonale, de la mélodie, de l'intensité, du débit et de l'intelligibilité de la parole pour évaluer la restitition d' une néoglotte et sa fonctionnalité.

Dans notre étude certains paramètres acoustiques s'améliorent franchement et rapidement, quand d'autres semblent plus lents à progresser. Les composants de la voix et de la parole dont l'évolution est la plus probante dans notre étude, sont *l'instabilité du voisement et l'intensité*. Ces résultats sont concordants avec ceux déja publiés (Arnoux-Sindt, 1992; Crevier-Buchman, 1995). L'intensité vocale varie avec la pression sous-glottique, et elle est liée à l'amplitude et à la nature du mouvement glottique. En effet, les structures néoglottiques dont dépend la voix, ont toutes été modifiées par la chirurgie: augmentation de la masse de la néo-glotte, de sa surface vibrante et de la tension des muscles sous-jacents. La sonorisation ne se fera qu'au prix de la mobilisation et de l'affrontement des

différents éléments anatomiques restants (Arnoux-Sindt, 1992). Sans une fermeture efficace à ce niveau, la pression sous-néoglottique ne sera pas suffisante pour donner de la puissance à la voix. La voix est de faible puissance et empreinte de désonorisations fréquentes à 1 mois post-opératoire mais, dès 6 mois, on observe un voisement constant et une intensité conversationnelle correcte pour 7 patients sur 10. L'évolution rapide de ces deux paramètres s'explique par la constitution d'un sphincter néo-glottique pour 7 patients sur 10 à 6 mois de leur intervention.

L'évolution du *timbre* rauque ou soufflé est moins franche et plus lente. Arnoux-Sindt (1992) fait cette observation en parlant d'un "aspect rapeux et guttural du timbre, qui devient progressivement plus souple". Il passe de très soufflé à 1 mois à voilé à 6 mois par amélioration de la qualité de fermeture de la néo-glotte (Crevier-Buchman, 1995).

La *hauteur* est un paramètre qui a été difficile à évaluer par le jury d'écoute. En effet, à 1 mois de l'intervention, certaines voix se sont révélées plus ou moins chuchotées d'où la difficulté d'évaluer leur hauteur. De plus, la composante de souffle et de bruit dans la voix ainsi que son caractère rauque n'ont pas facilité son évaluation.

A coté de l'ensemble des évolutions positives constatées, on remarque que l'adéquation de la *mélodie* au contexte ne suit pas cette progression et reste médiocre. Elle dépend de l'évolution temporelle de la fréquence fondamentale (FO) pendant l'émission vocale. La FO est déterminée par la vitesse de vibration de la muqueuse néo-glottique, en l'occurrence la muqueuse des aryténoïdes [Arnoux-Sindt, 1992; Crevier-Buchman, 1995). Cette muqueuse est de masse supérieure à celle des cordes vocales, et elle vibre sur une structure musculaire plus épaisse que celle des cordes vocales; la modification de la vitesse de vibration s'acquiert en général entre 18 et 24 mois post-opératoire (Crevier-Buchman, 1994 et 1995; Pech, 1988 et 1990).

Le *mécanisme vocal* fait intervenir les composantes d'*hypotonie* et d'*hypertonie* des muscles pharyngo-laryngés. On note une absence d'hypotonie dans la voix de nos patients. Par contre, il existe une légère hypertonie à l mois de l'intervention dans le but d'obtenir une meilleure occlusion néoglottique. L'évolution est favorable avec diminution progressive de ce comportement d'effort à 3 et 6 mois.

Le *débit* s'accélère nettement pour l'ensemble des patients. Cela montre l'acquisition progressive d'une bonne synergie pneumo-phonatoire.

La qualité de la voix et de la parole dans *l'expression spontanée par rapport à la lecture*, s'est révélée identique, pour la majorité des patients à 1, 3 et 6 mois. On peut en déduire que la gestion de leur comportement phonatoire, tant en situation contrainte de lecture qu'en expression libre est similaire.

L'évolution de l'*intelligibilité* est très nette, elle suit celle de l'intensité et du voisement.

Quant à *l'appréciation globale*, il existe une amélioration de la qualité de la voix et de la parole de ces patients. Ce dernier paramètre est en quelque sorte une synthèse de l'ensemble des composants de la voix et de la parole étudiés au cours de cette analyse. L'évolution constatée est moins significative entre 1 et 3 mois qu'entre 3 et 6 mois, comme le montrent les courbes concernant l'évolution du profil vocal « positif » (Tableau I).

CONCLUSIONS

On peut supposer que la néo-glotte se met en place et se mobilise pendant le premier trimestre avec l'acquisition du voisement et de l'intelligibilité, puis des paramètres prosodiques semblent s'installer en particulier l'intensité, la hauteur, et le débit phonatoire. Cependant, le comportement d'effort n'est pas tout à fait aboli. Ces constatations devraient améliorer la prise en charge rééducative de ces patients.

BIBLIOGRAPHIE

Piquet JJ, Desaulty A, Decroix G.(1974) Crico-hyoido-épiglotto-pexie. Technique opératoire et résultats fonctionnels. Ann Otolaryngol Chir Cervicofac;91:6819.

Piquet JJ, Darras JA, Berrier A, Roux X, Garcette L. (1986) Les laryngectomies subtotales fonctionnelles avec cricohyoidopexie. Technique, indication, résultats. Ann Oto-laryngol Chir Cervicofac; 103:411-5.

Laccourreye H, Laccourreye O, Weinstein G, Menard M, Brasnu D. (1990) Supracricoid laryngectomy with cricohyoidoepiglottopexy: a partial laryngeal procedure for glottic carcinoma. Ann Otol Rhinol Laryngol;99:421-6.

Guerrier B, Lallemant JG, Balmigere G, Bonnet Ph, Arnoux R. (1987) Notre expérience de la chirurgie reconstructive dans les cancers glottiques. Ann Otolaryngol Chir Cervicofac; 104: 175-9.

Crevier-Buchman L, Laccourreye O, Monfrais-Pfauwadel MC, Menard M, Jouffre V, Brasnu D. (1994) Evaluation informatisée des parametres acoustiques de la voix et de la parole apres laryngectomie partielle supra-cricoidienne avec cricohyoidoépiglotopexie. Ann Otolaryngol Chir Cervicofac; 3: 397-401.

Pech C, Giovani A, Henin N. (1990) La voix apres laryngectomies reconstructives. Bulletin d'audiophonologie; 6: 84107.

Pech C, Giovanni A, Henin N. (1988) La voix sans corde vocale. Rev Laryngol; 109:373-7.

Decroix G, Piquet JJ, Desaulty A, et al. (1976) Fonction vocale et laryngectomie reconstructive. Ann Otolaryngol Chir. Cervicofac; 93:129-48.

Traissac L. Réhabilitation de la voix et de la déglutition aprés chirurgie partielle ou totale du larynx. (1992) Rapport de la société française d'Oto-Rhino-Laryngologie et de pathologie cervico-faciale. Ed. Arnette, Paris, pp 227-38.

Bassich C, Ludlow C. (1986) The use of perceptual methods by new clinicians for assassing voice quality. J Speech Hear Disorders. 51: 125-33.

Fex S. (1992) Perceptual evaluation. J. Voice; 6: 155-8.

Hirano M. (1989) Objective evaluation of the human voice: clinical aspects. Folia Phoniatrica. 41: 89- 144.

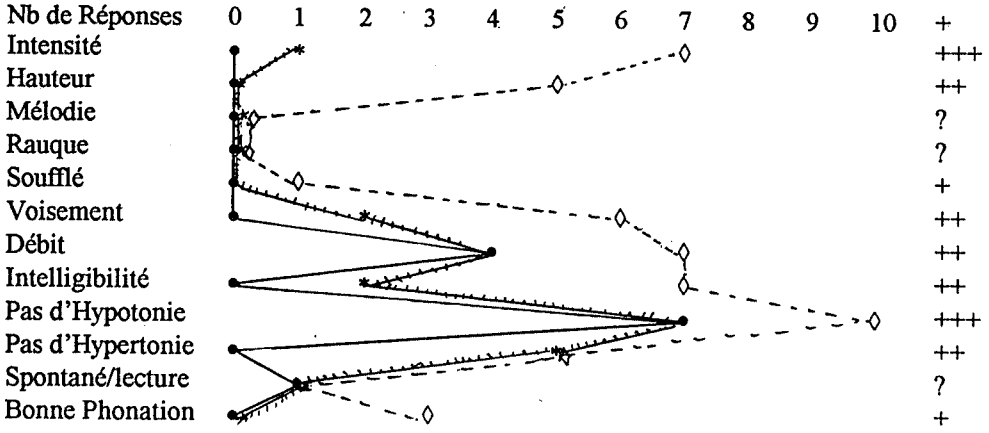
Dejonckere P. (1993) Perceptual evaluation of dysphonia: reliability and relevance. Folia Phoniatrica; 45: 76-83.

Arnoux-Sindt B. (1992) Articulation crico-aryténoidienne et laryngectomie reconstructive. Rev Laryngol. 113: 339-42.

Crevier-Buchman L, Laccourreye O, Weinstein G, Garcia D, Jouffre V, Brasnu D. (1995) Evolution of speech and voice following supracricoid partial laryngectomy. J Laryngol Otology . 109: 410-3.

Tableau I: Profil Vocal « Positif » des patients à 1 mois (•), 3 mois (*), et 6 mois (◊) post-opératoire.

Ce profil présente les réponses côtées O (Normal).



LIPCOM, UNE AIDE AUTOMATIQUE A LA LECTURE LABIALE

Audrey COURSANT-MOREAU, FRANCIS DESTOMBES
IBM France service 3090, 68/76 Quai de la Rapée; 75012 Paris
Tél.: (1) 40 01 53 22 - Fax: (1) 49 28 08 60 - e-mail: coursant@fr.ibm.com
Charlemagne s.a.r.l., Bellossy, 74160 Vers
Tél.: 50 35 25 44 - Fax: 50 35 25 52 - e-mail: 100730.663@compuserve.com

ABSTRACT

The IBM Lipcom is a research and development project; its aim is to help deaf persons to understand hearing people better via lip-reading. Lipcom uses a computer to perform real time, speaker-dependent phonetic recognition of continuous speech. The results are displayed on a computer screen in order to be perceived by the deaf person simultaneously with movements of the speaker's lips. This paper describes the system and the recognition technique used, then the experimentation phase in two schools for the deaf in Paris.

1. DESCRIPTION GÉNÉRALE

Le projet Lipcom est un effort de recherche et développement conduit par IBM France (Centre Scientifique). Il a pour but d'aider les personnes sourdes à mieux comprendre leurs interlocuteurs entendants au moyen de la lecture labiale. En effet, cette dernière est imprécise et fatigante, car beaucoup de sons de la parole ont une apparence identique sur les lèvres : par exemple, les syllabes (/pa/, /ba/, /ma/), (/fa/, /va/), etc.

Lipcom est un logiciel prototype qui effectue une reconnaissance phonétique de la parole continue en temps réel, en mode monolocuteur; des informations phonétiques sont affichées à l'écran en gros caractères, de façon à pouvoir être vues par les personnes sourdes en même temps que les mouvements des lèvres. La conjonction des deux sources d'information doit permettre une meilleure compréhension du message.

L'objectif de Lipcom est très voisin de celui du "Cued Speech" (Cornett 1967, Destombes 1982), qui utilise des "clefs phonétiques" manuelles pour fournir l'information manquante sur les lèvres. Lipcom n'est donc pas une aide à l'enseignement de la lecture labiale, mais un "outil prothétique" à utiliser conjointement avec cette dernière pour la rendre plus efficace.

L'expérimentation de Lipcom a commencé en mai 1994 dans une école pour enfants sourds à Paris, puis dans une deuxième fin 1995.

2. RECONNAISSANCE PHONÉTIQUE EN TEMPS RÉEL

2.1 Système et entrée acoustique

Lipcom utilise un ordinateur IBM PC sous DOS, et une carte audio IBM ACPA pour l'entrée et la sortie du signal vocal. Un ordinateur rapide (Pentium 120 MHz ou plus) est nécessaire pour effectuer les traitements acoustiques et la reconnaissance en temps réel.

Le signal vocal est échantillonné sur 16 bits à 14700 Hz, et traité par tranches de 128 échantillons, soit 8,7 ms. Les paramètres acoustiques calculés sont l'intensité de chaque tranche et le spectre de prédiction linéaire (LPC) sur les quatre dernières tranches saisies, soit 34,8 ms. La taille de cette fenêtre d'analyse est paramétrable, comme le nombre de coefficients LPC (ex.: 17).

2.2 Phonèmes, clés phonétiques, et clés graphiques

Lipcom essaie de reconnaître les phonèmes dans la parole continue avec le plus de précision possible. La plupart des phonèmes du français sont reconnus en tant que tels, mais dans certains cas, des regroupements sont nécessaires; par exemple, les phonèmes /b d g/ sont combinés en une "clé phonétique", qui apparaît à l'écran sous la forme d'une "clé graphique" (cf. Table 1).

Le choix des clés graphiques tient compte de critères tels que la facilité à les distinguer visuellement, ou la familiarité des enfants avec un système phonétique situé à mi-chemin entre l'A.P.I. et l'aphabet phonétique de S. Borel - Maisonny (S. Borel-Maisonny1955) employé en rééducation de la parole.

Table 1: Clés phonétiques et graphiques

clés phonétiques	clés graphiques	clés phonétiques	clés graphiques
/a/	A	/l/	L
/ɑ̃/	An	/R/	R
/o ɔ/	O	/f/	F
/ɔ̃/	On	/s/	S
/i/	I	/ʃ/	CH
/ɛ̃ œ̃/	In	/v/	V
/œ/	E	/z/	Z
/ø/	Eu	/ʒ/	J
/u/	Ou	/j/	I
/y/	U	/ɥ/	U
/é/	é	/w/	W
/ɛ/	è	/ptk/	★
		/bdg/	▲
		/mn/	MN

L A

Les clés vocaliques apparaissent dans la partie droite de l'écran, tandis que les clés consonantiques s'affichent dans la partie gauche, en jaune sur fond bleu (cf. table 1); ce mode de présentation a été défini à l'issue de tests de perception préliminaires.

2.3 Principes de reconnaissance

La reconnaissance phonétique s'effectue ainsi :

- Chaque tranche de signal est analysée, et ses paramètres acoustiques sont comparés à un grand nombre de modèles phonétiques, en utilisant une distance euclidienne.
- Les étiquettes phonétiques des deux plus proches voisins les plus récentes sont conservées dans une liste.
- Dans cette liste, les étiquettes successives identiques sont regroupées, et le calcul d'un score de reconnaissance décide de l'affichage ou non d'une clé graphique.

2.4 Définition et acquisition des modèles

Les modèles de phonèmes sont créés par segmentation et étiquetage manuels de phrases prononcées par le locuteur qui emploie le système. Cette tâche est aidée par l'affichage du signal vocal, des spectres correspondant aux extrémités d'un segment, ainsi que par des graphiques d'intensité et de variation spectrale.

Chaque modèle retenu correspond à une tranche de 8,7 ms à l'intérieur d'un segment phonémique; il est constitué de trois spectres : celui de la tranche (i) et ceux des tranches (i-k) et (i+k), ex.: k=4. Cet ensemble représente donc une sorte de "mini-spectrogramme local" pour chaque tranche d'un segment (Destombes 1995); elle offre l'avantage de modéliser l'évolution du spectre à l'intérieur d'un phonème et sa coarticulation avec les phonèmes qui l'entourent.

Le choix du corpus d'apprentissage a évidemment une grande importance; il faut de préférence utiliser des phrases comportant les digrammes et les trigammes les plus fréquents en français.

Bien que la création des modèles comporte une phase de "compactage", afin d'éliminer les modèles redondants trop proches les uns des autres, le nombre de modèles résultants est élevé. La recherche en temps réel des deux plus proches voisins a donc nécessité le développement de techniques de recherche rapide. Dans l'état actuel du prototype, la reconnaissance s'effectue en temps réel avec environ 15.000 modèles.

2.5 Visualisation des clés graphiques

Afin de permettre l'observation simultanée des lèvres du locuteur et des clés graphiques à l'écran, un dispositif très simple est utilisé : une vitre est placée à 45° entre le locuteur et les sujets sourds; par transparence, ces derniers peuvent voir le visage et les lèvres du locuteur; par réflexion, ils voient les informations de l'écran superposées au visage (cf. schéma 1).

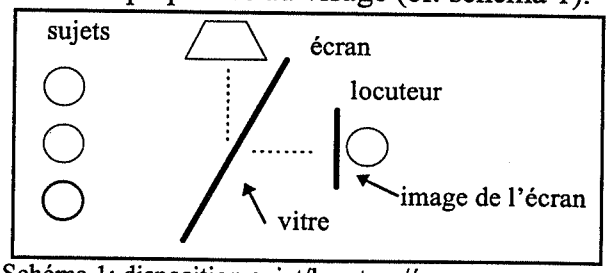


Schéma 1: disposition sujet/locuteur/écran.

2.6 Taux de reconnaissance

Le taux de reconnaissance dépend du type de corpus employé, mais il est en général supérieur à 95% pour les clés phonétiques (dans les corpus utilisés pendant les tests). Mais il se peut que des clés phonétiques supplémentaires apparaissent, par exemple lorsqu'un /ə/ final est confondu avec une voyelle, ou lorsqu'une semi-voyelle telle que /w/ conduit à l'affichage de "ow". Le taux des phonèmes ainsi rajoutés (ou omis) n'a pas encore été mesuré.

3. EXPÉRIMENTATION

3.1 Conditions

L'objectif de l'expérimentation menée depuis mars 1994 est d'évaluer l'utilité et l'efficacité de Lipcom en accordant une place considérable aux facteurs humains. Cette expérience est réalisée auprès de 11 enfants sourds profonds (âgés de 8 à 13 ans), dont le niveau scolaire est le CE2 pour les plus petits, et la 6ème pour les plus grands. La participation des enfants est d'une heure par semaine sur l'année scolaire.

Les premières expériences (Coursant 1995) ont permis de définir les modalités de présentation des clés à l'écran. Les séances de tests ont lieu dans une salle de classe ordinaire.

Le but de l'expérience décrite ici est d'évaluer l'apport de Lipcom dans une tâche d'identification de mots sosies labiaux, précédés de l'article défini. Le corpus est constitué de 120 noms communs de structures syllabiques CV, CVC, et CVCV. Pour éviter un effet de fatigue dû à la longueur du test, nous avons scindé le corpus en deux listes de 60 noms chacune (liste1 : CV+CVCV, liste2 : CVC). Chacun de ces noms a au moins un sosie labial dans la même liste, par exemple : *le riz, le lit,* ou *le mouton,le melon, le bouton...* Le choix du vocabulaire est adapté au registre lexical des sujets. Les conditions testées sont : lecture labiale + prothèse auditive (LL), puis lecture labiale + prothèse auditive + Lipcom (LLA, lecture labiale assistée).

Les sujets ont été entraînés pendant 10 séances (30 min. par semaine), à l'aide d'un programme de simulation de Lipcom ne faisant intervenir ni la reconnaissance de la parole, ni la lecture labiale. L'entraînement portait essentiellement sur l'apprentissage de la lecture des clés graphiques et de l'identification des mots. Quelques séances d'entraînement avec Lipcom ont également précédé les tests. L'éventail des mots utilisés pour lors des séances d'entraînement était beaucoup plus large que celui des tests, et comportait des mots de structures syllabiques très diverses.

3.2 Résultats

Les scores moyens d'identifications correctes sont de 48,5% en lecture labiale (LL) et de 76,5% en lecture labiale assistée (LLA), soit une amélioration de 28% sur l'ensemble des tests. Ces scores correspondent au nombre d'items (article défini + nom) parfaitement reconnus.

La valeur du test "t apparié" calculée sur l'ensemble des scores est de 5,681 avec $p>0,0013$; les résultats obtenus en condition LLA sont donc significativement meilleurs que ceux obtenus en condition LL.

Les écarts types relevés pour ces deux conditions, 15,4% en LL et 9,8% en LLA, révèlent des différences importantes de scores entre les sujets, qui s'atténuent avec l'utilisation de Lipcom. Les raisons qui pourraient expliquer ces écarts sont diverses :

- le gain apporté par la prothèse auditive diffère d'un sujet à l'autre,
- l'aptitude à lire sur les lèvres n'est pas identique pour tous, (Gentil 1981)
- les performances dans l'utilisation de codes phonologiques pour l'identification de mots peuvent elles-mêmes dépendre du degré d'intelligibilité de la parole de la personne sourde (Conrad 1970, 72, 79), de la fréquence d'occurrence des mots rapportée au registre lexical des sujets, et de la complexité de la correspondance phonie-graphie (Dodd 1987, Leybaert & Alégria 1991).
- en LLA les scores de Lipcom varient d'un test à l'autre.

Les confusions relevées sur les tests en LL représentent 36,5% des réponses : 27% correspondent strictement à des sosies labiaux du stimulus et 7,5% sont des mots inventés par les sujets mais qui ont le plus souvent la même image labiale que le stimulus (ex.: *soud pour sud, vousez pour fusée...).*

En LLA les 23,5% d'erreurs des sujets se ventilent de la façon suivante :

- 10% : confusions avec sosies labiaux
- 4% : confusions inexpliquées
- 9,5% : absences de réponse

Le taux de confusion avec des sosies labiaux montre que les sujets privilégient l'information visible sur les lèvres; en ce sens ils respectent la consigne de départ, "regarder les lèvres".

Par ailleurs, sur les 23,5% d'erreurs, seuls 8% peuvent être attribués aux erreurs de

reconnaissance de Lipcom. Les scores du système (taux de reconnaissance des items) varient entre 63,5% et 90%. Il n'y a pas de corrélation apparente entre ces scores et ceux obtenus par les sujets, comme le montre le tableau ci-après :

Table 2: Scores d'identification obtenus par les sujets - scores de reconnaissance de Lipcom pour les différents tests en LLA.

Liste de mots	Groupes de sujets	Scores des sujets	Scores de Lipcom[1]
CV + CVCV	grands	79%	items: 86,5% phon: 98%
	petits	89%	items: 73% phon: 93%
CVC	grands	70%	items: 90% phon: 98%
	petits	66,5%	items: 63,5% phon: 95%

Il est important de souligner que toutes les erreurs de Lipcom n'entraînent pas obligatoirement des réponses erronées chez les sujets; tout dépend en effet du type d'erreur dans la reconnaissance (omission, insertion ou confusion de phonème), mais également de la place de cette erreur dans l'item.

Les résultats obtenus par les sujets sont plus élevés pour la liste CV+CVCV que pour la liste CVC; notre hypothèse est que le mode d'affichage des clés graphiques (consonnes à gauche sur l'écran et voyelles à droite) a pu influencer ces résultats. Certains sujets avaient l'impression que le temps d'apparition des clésgraphiques était plus court pour la deuxième liste que pour la première.

Par ailleurs, il est vraisemblable que les enfants sont influencés par la fréquence d'utilisation des mots dans leur entourage. Par exemple, pour le stimulus "pouce" ils ont souvent répondu par le sosie labial "bus", qui leur est plus familier, même si la reconnaissance de Lipcom était exacte.

Enfin, il est ressorti clairement que l'accumulation de clés génériques telles que ▲ (/b d g/) et ★ (/p t k/), au sein d'un même mot, retardait souvent son identification.

[1] On distingue ici le % d'items (nom + article) parfaitement reconnus, du % global de phonèmes reconnus.

4. CONCLUSION

Dans cette expérience, l'utilisation de Lipcom permet d'améliorer de façon significative la compréhension en lecture labiale assistée avec 76,5% d'identifications correctes contre 48,5% en lecture labiale seule. Ces résultats sont d'autant plus encourageants que le système est encore très imparfait. L'amélioration de la reconnaissance fait partie des objectifs prioritaires. Par ailleurs, cette expérience a mis l'accent sur des difficultés liées au mode d'affichage des informations; d'autres solutions devront être envisagées dans la perspective d'utiliser des corpus plus complexes.

BIBLIOGRAPHIE

BOREL-MAISONNY S. (1955), *Langage oral et écrit I, Delachaux et Niestlé, Neuchatel.*

CONRAD R. (1970), Short-term memory processes in the deaf, *British Journal of Psychology, 61, 179-195.*

CONRAD R. (1979), *The deaf school child, London, Harper & Row.*

CONRAD R. (1972), Speech and reading, *in KAVANAGH & MATTINGLY (Eds), Language by ear and by eye, Cambridge, Mass., MIT Press.*

CORNETT R.O. (1967), Cued Speech, *American Annals of the Deaf, 112, 3-13.*

COURSANT A. (1995), Une aide automatique à la lecture labiale pour les sourds : Lipcom (IBM), Premières expériences, *Travaux de l'Institut de Phonétique de Strasbourg, n°25, 1-19.*

DESTOMBES F. (1982), *Aides manuelles à la lecture labiale et perspectives d'aides automatiques, Étude f054 Centre Scientifique IBM France.*

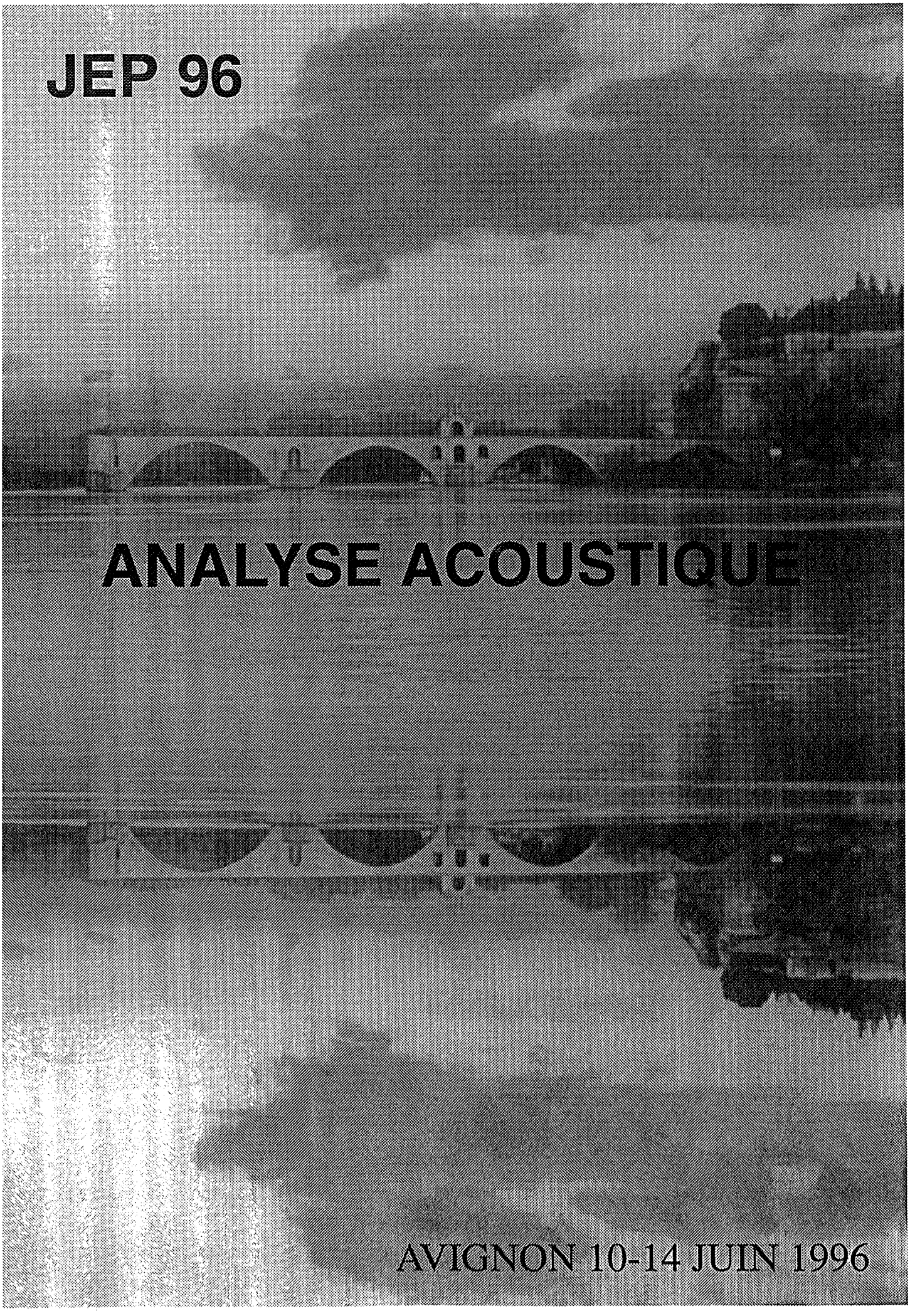
DESTOMBES F. (1995), *Phonetic speech recognition using spectrum pairs, IBM Technical Disclosure Bulletin, Vol 38 N° 04.*

DODD B. (1987), *Lipreading, phonological coding and deafness, in B. DODD & R. CAMPBELL (Eds.) Hearing by eye : the psychology of lipreading, Lawrence Erlbaum Associates, London.*

GENTIL M. (1981), Étude de la perception de la parole : lecture labiale et sosies labiaux, Centre Scientifique IBM France.

LEYBAERT J. & ALEGRIA J. (1991), *Mécanisme d'identification des mots chez le sourd.* La reconnaissance des mots dans les différentes modalités sensorielles. Psychologie d'aujourd'hui, PUF.

JEP 96



ANALYSE ACOUSTIQUE

AVIGNON 10-14 JUIN 1996

REPRESENTATION DU SPECTRE DE PAROLE PAR LES MULTIGRAMMES

Jan Cernocky [1,2], Geneviève Baudoin [1]

(1) ESIEE, Département Signal, Cité Descartes BP 99, 93162 Noisy-le-Grand Cedex,
tél: 45.92.66.91, 45.92.66.46, fax: 45.92.66.99, email: cernockj@esiee.fr, baudoing@esiee.fr
(2) FEI VUT, Institut de la Radioélectronique, Antoninska 1, 662 09 Brno, République Tchèque,
tél: +42-5-41.32.12.75, fax: +42-5-41.21.11.35, email: cernocky@ant.fee.vutbr.cz

ABSTRACT

The multigram segmentation allows an efficient representation of speech spectrum (decrease of information amount for a given spectral distortion). First, the standard multigram segmentation was applied to vector quantized cepstral vectors. The results, except for small QV codebooks, are not satisfactory. Second, a modification of the method was developed using a multigram-distance notion. The segmentation and dictionary training procedure are presented, as well as experimental results.

1. INTRODUCTION

Dans les différents domaines du traitement numérique de la parole, il est nécessaire de représenter efficacement l'enveloppe spectrale des trames de parole. On utilise généralement la quantification vectorielle (QV) des vecteurs de coefficients spectraux pour réduire le nombre de bits nécessaire. Cependant, la QV n'opère que dans le domaine spatial (elle partitionne efficacement l'espace des vecteurs), et elle ne prend pas en compte les dépendances existant entre les trames. De nombreuses approches exploitant ces dépendances et appelées *techniques segmentales* ont été récemment décrites. On peut diviser les techniques employées en *Quantification Matricielle* (Shiraki, 1988, Chou, 1994), *Codage Multi-trame* (Lopez, 1993) et *Codage avec Segmentaion Phonétique* (Wang, 1991). Le problème principal de ces techniques est le choix des segments.

Dans le travail, on a utilisé la technique des *multigrammes* qu'on a appliqué d'abord à la segmentation des chaînes de symboles provenant de la quantification vectorielle. On montre, que l'utilisation directe de cette technique n'a pas d'intérêt par rapport à la QV simple: il est impossible de trouver des séquences caractéristiques de symboles, surtout pour des grands codebooks de QV. On a développé une extension de cette méthode, qu'on appelle *multigrammes avec distance* ou *multigrammes modifiés*. Elle fournit de nettement meilleurs résultats en termes de taille du dictionnaire et de réduction du nombre de bits nécessaire pour représenter le spectre en gardant la même distortion spectrale moyenne.

Remarque: Dans cet article, on utilise le terme anglais *codebook* pour le dictionnaire de QV et on réserve le mot *dictionnaire* pour l'ensemble des séquences caractéristiques de longueur variable (multigrammes).

2. MULTIGRAMMES

Les articles de Bimbot et al. (Bimbot, 1994; Deligne, 1995) présentent cette méthode en détail.

2.1. Segmentation, dictionnaire

Une chaine de symboles de longueur N, $W=w_1w_2...w_N$ est segmentée en une suite de séquences $S_1S_2...S_q$ dont la longueur peut varier de 1 à n (le mot *multi*gramme provient de cette variabilité). La segmentation optimale parmi l'ensemble de toutes les segmentation possibles $\{B\}$ est choisie afin de maximiser la vraisemblance:

$$L(W) = \max_{\{B\}} \prod_{k=1}^{q} p(S_k) \qquad (1)$$

Comme les probabilités des séquences S_k ne sont pas connues a-priori, on crée un *dictionnaire* de séquences caractéristiques à l'aide d'une chaîne d'apprentissage. Les probabilités sont initialisées en prenant en compte toutes les séquences apparaissant dans la chaîne, puis réestimées dans un processus itératif à deux étapes:

- *segmentation*, où la chaîne est segmentée en utilisant l'Equation 1,
- *réestimation des probabilités* par:

La recherche a été soutenue par la bourse du Gouvernement Français No. 94/4516 et par le Programme Elite de Texas Instruments.

$$p(S) = \frac{c(S)}{C} \quad (2)$$

où $c(S)$ est le nombre d'occurences de séquences S dans la chaîne segmentée et C le nombre total de séquences dans la segmentation. La probabilité peut également être réestimée en utilisant un facteur de pénalisation a (qui aide à éliminer les séquences rares du dictionnaire):

$$p_a(S) = \frac{c(S)}{C}\left(1 - a\sqrt{\frac{C - c(S)}{c(S).C}}\right) \quad (3)$$

Notons, que lors de la dernière itération, les probabilités doivent être obligatoirement réestimées en utilisant l'Equation 2.

2.2. Evaluation

Pour évaluer l'intérêt de la segmentation en multigrammes par rapport à la QV, on compare deux entropies: d'une part celle du codebook de la QV, calculée lors de l'apprentissage de la QV:

$$H(V) = -\sum_{i=1}^{L} p(y_i)\log_2 p(y_i) \quad (4)$$

où y_i sont les vecteurs-code et L est leur nombre, d'autre part l'entropie par symbole du dictionnaire des multigrammes:

$$H'(M) = \frac{-\sum_{i=1}^{Z} p(M_i)\log_2 p(M_i)}{\sum_{i=1}^{Z} l(M_i)p(M_i)} \quad (5)$$

où Z est le nombre de multigrammes dans le dictionnaire, $p(M_i)$ est la probabilité et $l(M_i)$ la longueur du multigramme M_i. Une fois évaluées sur la chaîne d'apprentissage, on valide ces entropies sur une *chaîne de test* en estimant le nombre de bits moyen nécessaire pour représenter un vecteur spectral (le *débit* en terminologie du codage):

$$R(V) = -\sum_{i=1}^{L} c_{test}(y_i)\log_2 p(y_i) \,/\, N_{test} \quad (6)$$

pour la QV, où $c_{test}(y_i)$ est le nombre de vecteurs de la chaîne de test, représentés par le vecteur-code y_i et

$$R(M) = \frac{-\sum_{i=1}^{Z} c_{test}(M_i)\log_2 p(M_i)}{N_{test}} \quad (7)$$

pour les multigrammes, où $c_{test}(M_i)$ est le nombre de séquences représentées par le multigramme M_i. Dans les deux cas, N_{test} est la longueur de la chaîne de test.

2.3. Expériences et résultats

Pour les expériences, on a utilisé une base de données monolocuteur, d'où on a tiré 213270 trames de parole pour l'apprentissage et 122903 trames pour les tests (longueur de trames 20 ms, récouvrement 10 ms). Les spectres ont été représentés par 10 coefficients cepstraux calculés à partir de 10 coefficients LPC. On a effectué l'apprentissage des codebooks de QV et la quantification, dont les résultats sont donnée dans la Table 1.

Table 1. Résultats de la QV: taille du codebook, entropie, débit pour la chaîne de test, distortion spectrale (distance Euclidienne sur les c_i) moyenne.

L	H(V) [bit]	R(V) [bit]	SD [dB]
2	0.999	1.000	4.332
4	1.950	1.930	3.770
8	2.914	2.888	3.199
16	3.881	3.844	2.893
32	4.863	4.830	2.640
64	5.869	5.848	2.413
128	6.858	6.840	2.231
256	7.865	7.853	2.055
512	8.865	8.852	1.907

Les apprentissages et les tests des multigrammes ont été effectués pour des longueurs maximales de n=2, 4, 6, 8, 10 et pour des facteurs de pénalisation de a=0, 0.5, 1, 1.5, 2, 2.5, 3. Les résultats optimaux pour deux critères différents: entropie par symbole minimale du dictionnaire (1), débit minimal pour la chaîne de test (2), sont donnés pour deux codebooks de QV: L=16 et L=128 (Table 2).

Table 2. Multigrammes classiques: n optimal, a optimal, taille du dictionnaire, entropie par symbole du dictionnaire, débit pour la chaîne de test.

L	16(1)	**16(2)**	128(1)	**128(2)**
n_{opt}	10	**6**	10	**2**
a_{opt}	0.5	**1.0**	0.0	**0.0**
taille	16225	**2593**	21368	**3451**
H'(M)	1.407	**2.264**	1.453	**5.414**
R(M)	4.051	**2.328**	12.737	**5.465**

En ne regardant que les résultats optimaux lors de l'apprentissage, on constate une baisse significative de l'entropie par rapport à la QV correspondante. Malheureusement, cette décroissance est due à un fort sur-apprentissage, la taille du dictionnaire et le débit obtenu sur la chaîne de test en sont la preuve. Pour la segmentation d'autres chaînes que celle d'apprentissage (ce qui est normalement le cas), les résultats optimaux sont donnés par le débit le plus petit obtenu

sur la chaîne de test (imprimés en gras dans la Table 2). On peut ainsi économiser 1.61 bit (41%) pour L=16 et 1.44 bit (21%) au cas L=128 par rapport aux QVs correspondantes.

En regardant les longueurs maximales optimales (2 pour L=128), on peut conclure, qu'il est difficile de trouver les séquences de symboles caractéristiques pour des plus grands codebooks. Les petits codebooks, où c'est possible, ne sont pas assez précis en ce qui concerne la distortion spectrale moyenne.

3. MULTIGRAMMES MODIFIES

Cette difficulté de trouver les séquences caractéristiques s'explique par la nature de la méthode des multigrammes: celle-ci exige l'identité stricte entre une séquence et un multigramme dans le dictionnaire. Dans le traitement des spectres, on peut relâcher cette contrainte à condition que la distortion spectrale moyenne ne soit pas dégradée. On définit une *distance* entre deux suites de vecteurs comme la moyenne de leurs distances Euclidiennes:

$$D(X,Y)=\frac{1}{l}\sum_{i=1}^{l} d(x_i,y_i) \qquad (8)$$

Dans notre cas, les longueurs de deux séquences sont égales, mais une comparaison de deux séquences de longueurs inégales est aussi envisageable, avec un alignement temporel.

Avec la notion de distance, on définit une modification de la méthode citée plus haut, qu'on appelle *multigrammes modifiés*. On cherche à diviser une chaîne en une suite de séquences, mais cette fois ci, la chaîne n'est plus constituée des symboles, mais de vecteurs: $X=x_1x_2...x_N$, où, dans notre cas, les x_i sont les vecteurs cepstraux. De même, les multigrammes dans le dictionnaire, ne sont plus des séquences de symboles, mais des séquences de vecteurs: $M_i=m_{i,1},m_{i,2}...m_{i,li}$ où l_i est la longueur de M_i. On les appelle *multigramme-codes*.

3.1. Segmentation, dictionnaire

La maximisation de la quantité $L'(X)$, similaire à l'Equation 1, est utilisée pour la segmentation de la chaîne X en une suite de séquences $U_1U_2...U_q$:

$$L'(X)=\max_{\{B\}}\prod_{k=1}^{q} p'(M_k) \qquad (8)$$

avec $p'(M_k)=Q[D(U_k,M_k)].p(M_k)$ où $p(M_k)$ est la probabilité du multigramme-code M_k qui représente la séquence U_k. Le multigramme optimal est choisi afin de minimiser la distance $D(U_k, M_k)$. Sa *probabilité pénalisée* $p'(M_k)$ doit décroître lorsque la distance D augmente; on applique une fonction de pénalisation Q donnée par:

$$Q[D]=\begin{cases}1-D/D_{\max} & pour \quad D\le D_{\max}\\ 0 & pour \quad D>D_{\max}\end{cases} \qquad (10)$$

où D_{max} est une distance maximale.

Le dictionnaire contient les vecteurs des multigramme-codes et leurs probabilités. Il est également créé par un processus itératif. Après l'initialisation (pour cela, on utilise les multigrammes classiques les plus probables et un codebook de QV), une itération se compose de trois étapes:

- *segmentation* de la chaîne, par maximisation de $L'(X)$.
- *réestimation des probabilités,* donnée pour le multigramme-code M_i par $p(M_i)=c(M_i)/C$, où $c(M_i)$ est le nombre de séquences représentées par M_i lors de la segmentation et C est le nombre total de séquences.
- *recalcul des multigramme-codes*, qui sont donnés par les centroïdes des séquences représentées par le multigramme correspondant.

3.2. Evaluation

Les résultats sont évalués en termes de trois paramètres: la taille du dictionnaire des multigrammes modifiés, la distortion spectrale moyenne et l'entropie par symbole (voir l'Equation 5) du dictionnaire. La dernière est vérifiée par l'évaluation du débit sur une chaîne de test (l'Equation 7).

3.3. Résultats expérimentaux

Le problème crucial pour les multigrammes modifiés est le choix de leurs nombres initiaux. Or, ces nombres peuvent décroître lors des itérations, mais on ne peut pas créer de nouveaux multigrammes. On a initialisé le dictionnaire avec plusieurs codebooks de QV (taille L=2, 4, 8, 16, 32, 64, 128, 256, 512) et avec les dictionnaires des multigrammes classiques pré-calculés. On a limité les nombres de multigrammes à L pour les 1-grammes et à $2L$ pour les 2- à n-grammes. On a travaillé avec les multigrammes de longueur maximale n=5, et 5 itérations dans l'apprentissage. Pour chaque initialisation, on a évalué les

résultats pour plusieurs distances maximales D_{max} (voir l'Equation 10).

Les résultats pour toutes les initialisations sont regroupés dans la Figure 1, où l'on peut les comparer à la courbe distortion-entropie de la QV. Les résultats détaillés pour l'initialisation L=128, sont donnés dans la Table 3.

Table 3. Résultats pour les multigrammes modifiés avec l'initialisation L=128. Distance maximale, distortion spectrale moyenne, nombres de 1- à 5-grammes dans le dictionnaire, entropie par symbole, débit pour la chaîne de test. * - échec de la segmentation dû au nombre de 1-grammes 0.

D_{max}	SD [dB]	1	2	3	4	5	H'(M) [bit]	R(M) [bit]
0.1	2.209	128	0	0	0	0	6.867	6.842
0.2	2.203	128	165	70	5	2	6.888	6.865
0.3	2.170	128	255	244	195	185	5.966	6.018
0.4	2.245	128	241	249	252	252	3.602	3.595
0.5	2.472	127	191	217	234	255	2.068	2.060
0.6	2.666	71	88	131	151	256	1.539	1.520
0.7	2.750	16	13	33	67	256	1.429	1.403
0.8	2.768	2	0	11	35	256	1.414	1.381
0.9	2.778	0	0	2	30	255	1.398	*
1.0	2.784	0	0	2	31	256	1.391	1.525

On peut conclure, que les multigrammes modifiés offrent, pour la représentation des spectres, de nettement meilleurs résultats que les multigrammes classiques. On a besoin de significativement moins de bits pour représenter un vecteur spectral qu'avec l'utilisation de la QV, tout en gardant une distortion spectrale comparable. Cependant, la méthode n'est pas complétement sans problèmes:

- Le choix de la distance maximale D_{max} est critique pour le rapport distortion-débit. Par contre, cette dépendance pourrait être utilisée dans les codeurs à débit variable.
- la complexité du calcul est importante.

3.4. Comparaison avec la VVVQ

La méthode de multigrammes modifiés peut être comparée à la Quantification vectorielle de séquences de longueur variable avec des symboles de longueur variable (Variable to variable length vector quantization) proposée par (Chou, 1994). Notre approche consiste en une généralisation des multigrammes. La méthode proposée par Chou est obtenue en généralisant le principe de la Quantification vectorielle avec contrainte entropique (ECVQ) à la Quantification matricielle (MQ).

4. CONCLUSION

L'utilisation directe des multigrammes classiques sur les vecteurs quantifiés vectoriellement n'apporte pas de bons résultats. Par contre, les multigrammes modifiés nous fournissent une représentation efficace du spectre par rapport à la QV. Dans le domaine du codage, on peut envisager l'application de cette méthode aux sous-vecteurs spectraux (split-quantization), afin d'obtenir une qualité transparente avec une complexité raisonable. En reconnaissance, les multigrammes modifiés moins précis et avec un alignement temporel peuvent devenir des unités efficaces pour le décodage acoustico-phonétique.

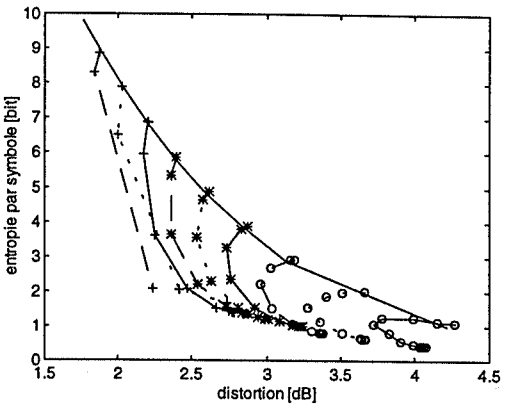


Figure 1. Multigrammes modifiés: résultats pour les initialisations L=2(-o-), 4(...o...), 8(---o---), 16(-*-), 32(...*...), 64(---*---), 128(-+-), 256(...+...), 512(---+---), comparaison avec la QV (sans marqueurs).

5. BIBLIOGRAPHIE

Shiraki Y., Honda M. (1988) LPC speech coding based on variable length segment quantization, *IEEE Trans. on ASSP*, Vol 36, No. 9, pp. 1437-1444.

Lopez-Soler J.M., Farvardin N. (1993) A combined quantization-interpolation scheme for very low bit rate coding of speech LSP parameters, *ICASSP 93*, pp. II-21-24.

Wang S., Gersho A. (1991) Phonetic segmentation for very low rate speech coding, *Advances in speech coding*, Kluwer Publishers, pp. 225-234.

Bimbot F., Pieraccini R., Levin E., Atal B. (1994) Modèles de séquences à horizon variable: Multigrams, *JEP 94*, pp. 467-472.

Deligne S., Bimbot F. (1995) Language modelling by variable length sequences: Theoretical formulation and evaluation of multigrams, *ICASSP 95*, pp. 169-172.

Chou P.A., Lookabaugh T. (1994) Variable dimension vector quantization of linear predictive coefficients of speech, *ICASSP 94*, pp. I-501-508.

ETUDE ET ANALYSE PAR LA METHODE TLM DE LA PROPAGATION ACOUSTIQUE DANS LE CONDUIT VOCAL : EFFET DES MODES D'ORDRE SUPERIEUR

Samir El-Masri (1, 2), Xavier Pelorson (1), Pierre Saguet (2), Pierre Badin (1)

(1) Institut de la Communication Parlée URA CNRS 368, INPG - UNIV. Stendhal
46, Av. Félix Viallet 38031 Grenoble CEDEX 1 FRANCE

(2) Laboratoire d'Electromagnétisme, Micro-ondes et Optoélectronique, 23 Rue des Martyrs, B.P. 257, 38016 Grenoble Cedex FRANCE. Tél.: 76 57 45 34 - Fax: 76 57 47 10 - e-mail: elmasri@icp.grenet.fr

ABSTRACT

This paper describes the application of the numerical method TLM (*Transmission Line Matrix*) to the solving of acoustical propagation problems. In particular, the problem of boundary conditions for the radiation into the infinite space has been dealt with, and the TLM method has been validated on simple geometrical cases, with various types of elementary sources. Higher order propagation modes have also been studied, and pressure reflection coefficients at the end of a rectangular guide have been determined separately for each of the plane and transverse modes. The encouraging results obtained allow to envisage the application of the method to more complex vocal tracts.

1. INTRODUCTION

La plupart des théories actuelles sur la propagation des ondes sonores dans le conduit vocal reposent sur une description unidimensionnelle qui n'est valable qu'aux basses fréquences (Fant, 1960). L'étude des phénomènes intervenant aux hautes fréquences sont difficiles à quantifier, non seulement d'un point de vue théorique mais aussi expérimental. Les méthodes de simulations numériques sont donc des outils très populaires (Lu et al., 1993; Matsuzaki et al., 1995).

Dans cet article nous présentons tout d'abord le principe de la méthode TLM appliquée à l'étude des ondes acoustiques. Cette méthode, utilisée jusqu'ici dans le domaine de l'électromagnétisme (Johns et al., 1971; Saguet, 1985), peut être appliquée aux ondes acoustiques moyennant des modifications du processus. Comparée aux méthodes par éléments finis, cette méthode de simulation numérique en trois dimensions est caractérisée par sa simplicité et sa facilité d'implémentation. Nous présentons ensuite quelques exemples de validation de la méthode. Nous décrivons enfin une étude de l'influence des modes d'ordre supérieur sur la propagation et le rayonnement des ondes sonores pour des modèles simplifiés du conduit vocal. Les résultats obtenus montrent clairement l'importance des modes transverses, qui ne peuvent être décrits par les modèles théoriques unidimensionnels.

2. PRINCIPE DE LA METHODE TLM

La méthode TLM a été élaborée au début des années soixante-dix par Johns & Beurle (1971), à partir de considérations sur le principe de Huygens en optique que l'on peut résumer ainsi: *"Chaque point d'un front d'onde peut être considéré comme étant le centre d'une perturbation qui donne naissance à des ondes secondaires sphériques. Le front d'onde est l'enveloppe de ces ondes sphériques"*

Ce principe a été appliqué à la résolution de problèmes d'électromagnétisme en créant les sources secondaires par la connexion de lignes de transmission régulièrement disposées suivant les axes de la structure étudiée : en appliquant une impulsion de Dirac en un noeud (point d'excitation) nous obtenons des impulsions qui se propagent en un temps Δt vers les noeuds voisins (voir Figure 1). Les impulsions se répartissent sur les différents bras des noeuds et se propagent à nouveau, ce qui permet de simuler la propagation des ondes électromagnétiques dans le domaine temporel.

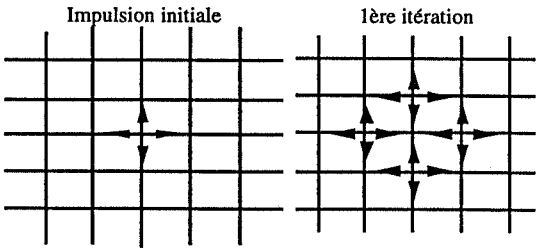


Figure 1 : Propagation des impulsions dans un milieu à deux dimensions (d'après Saguet, 1985).

2.1. Justification analytique en 3 dimensions de la méthode en acoustique

Le schéma équivalent électrique du noeud élémentaire de base (en acoustique, noeud scalaire) se présente alors sous la forme donnée à la Figure 2 (Hoefer, 1989): L et C sont respectivement l'inductance et la capacitance linéiques, Δl est la distance entre deux noeuds consécutifs, V_y est la tension du noeud par rapport à la masse. La répartition des tensions et des courants au noeud donne l'équation d'onde suivante :

$$\Delta V_y = 3LC\frac{\partial^2 V_y}{\partial t^2}$$

Cette équation présente une analogie directe avec l'équation d'onde acoustique :

$$\Delta P = \frac{1}{c^2}\frac{\partial^2 P}{\partial t^2}$$

où P est la pression et c la vitesse du son.

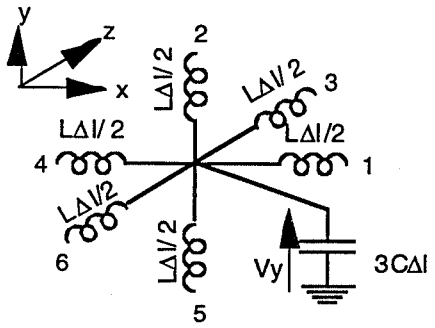


Figure 2 : Schéma électrique d'un noeud scalaire en trois dimensions

Il apparaît donc qu'il y a équivalence entre la tension V_y et la pression acoustique P d'une part, et entre les courants I_x, I_y, I_z et les vitesses acoustiques v_x, v_y, v_z suivant les 3 axes d'autre part.

2.2. Conditions aux limites

Les parois sont supposées se trouver à égale distance entre deux noeuds, afin que l'impulsion réfléchie revienne en phase sur le noeud d'origine.

Dans le cas de conditions aux limites parfaitement réfléchissantes, la vitesse normale à une paroi est nécessairement nulle. Ceci correspond sur la ligne de transmission équivalente à un coefficient de réflexion des impulsions égal à 1.

Pour des structures ouvertes sur l'espace infini, la réflexion des impulsions peut être calculée à l'aide de l'équation de prédiction linéaire suivante (Saguet, 1991):

$$V_{yr}(N_y,t) = 2.5V_{yr}(N_y-1,t-1) - 2V_{yr}(N_y-2,t-2) + 0.5V_{yr}(N_y-3,t-3)$$

où $V_{yr}(N_y,t)$ est le coefficient de réflexion à l'instant t pour le noeud N_y suivant l'axe y (voir Figure 3).

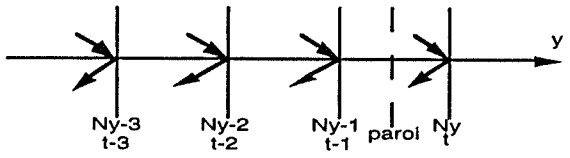


Figure 3 : réflexions aux parois, structures ouvertes

Bien entendu, cette méthode implique un maillage du voisinage immédiat de l'extrémité du guide.

3. VALIDATION DE LA METHODE TLM

Dans la suite, afin de valider et d'évaluer la précision de la simulation numérique, nous présentons une comparaison entre les résultats obtenus par la méthode TLM et les calculs théoriques.

3.1. Modélisation des sources

Nous étudions tout d'abord le rayonnement de sources élémentaires dans l'espace infini. La Figure 4 présente une comparaison entre les résultats théoriques et les résultats obtenus par la méthode TLM pour deux types de sources élémentaires : le monopole et le quadripôle latéral.

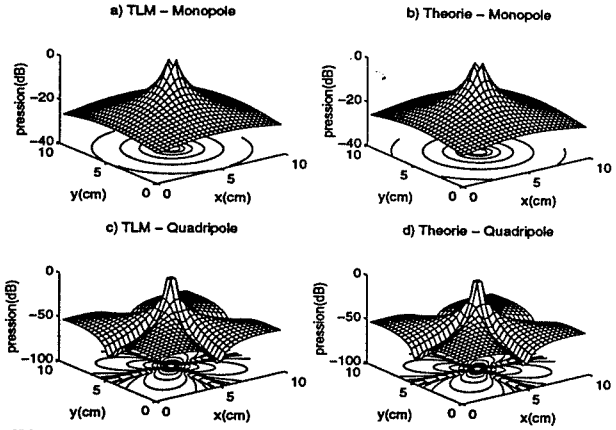


Figure 4 : Comparaison entre théorie et TLM pour les champs de pression rayonnés par une source monopolaire et une source quadripôlaire dans l'espace infini (fréquence de source: 9 kHz).

On constate un bon accord entre théorie et simulation, avec une différence maximale de ±1.5 dB au voisinage de la source. Une précision meilleure pourrait être obtenue au moyen d'un maillage plus fin, au détriment du temps de calcul.

3.2. Fonction de transfert d'un tuyau ouvert aux basses fréquences

Nous considérons maintenant le cas d'un conduit acoustique rectangulaire, uniforme et débouchant dans un écran plan infini. Le maillage de cette structure est présenté à la Figure 5.

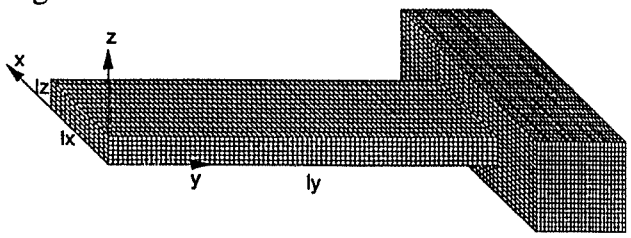


Figure 5 : Maillage du conduit rectangulaire uniforme.

Pour cette simulation, la source monopolaire est située dans un coin (x = 0.125 cm, y = 0.125 cm, z = 0.125 cm), et le récepteur se trouve à l'extrémité du guide sur l'axe médian (x = 1.375 cm, y = 20 cm, z = 0.625 cm). La Figure 6, qui présente une comparaison entre les fonctions de transfert théoriques et simulées, montre que

l'accord entre la théorie (Bruneau, 1983) et la simulation est excellent.

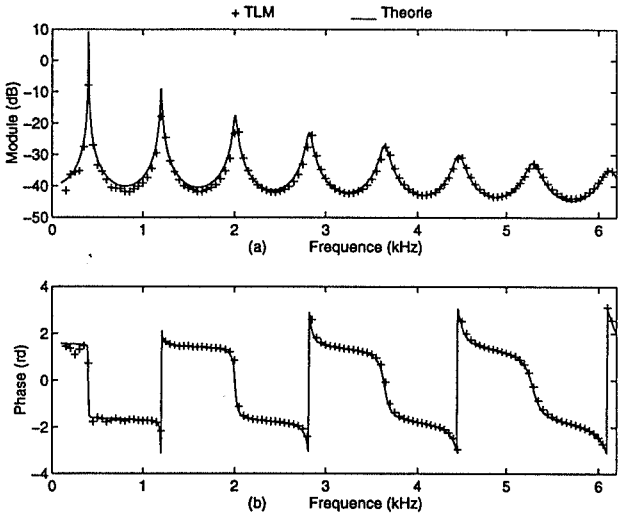


Figure 6 : Comparaison entre la fonction de transfert théorique (—) et la simulation numérique (+).

4. ETUDE DES MODES SUPERIEURS

4.1. Mise en évidence

D'un point de vue théorique, il est bien connu que l'hypothèse de propagation unidimensionnelle (ondes planes) n'est valable qu'en dessous de la première fréquence de coupure des modes d'ordre supérieur. L'existence et l'importance de cette fréquence de coupure ont été mises en évidence expérimentalement par Pelorson et al. (1995). Ces résultats sont confirmés par nos simulations numériques comme montré à la Figure 7.

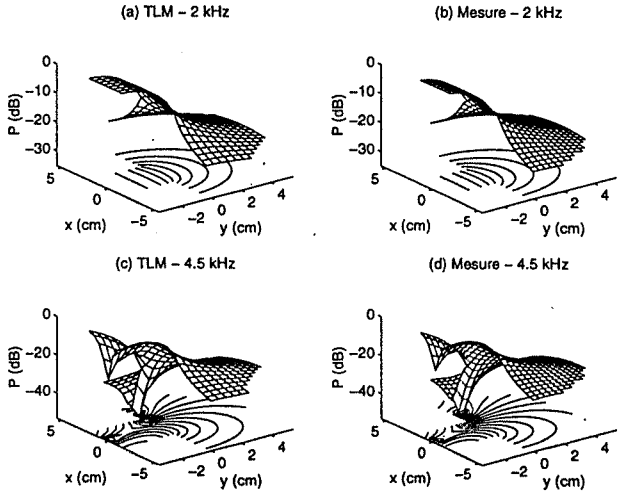


Figure 7 : Champ de pression à l'intérieur et à l'extérieur d'un conduit rectangulaire baffle. A droite : résultats de la simulation TLM à 2 kHz et 4.5 kHz; à gauche: résultats des mesures aux mêmes fréquences.

La Figure 7 représente le champ de pression mesuré et simulé à l'intérieur et à l'extérieur d'un tuyau rectangulaire de dimensions (4 cm × 60 cm × 1.5 cm). Le détail de la procédure expérimentale est décrit par Pelorson et al. (1995). On constate qu'aux basses fréquences (en dessous de 4.3 kHz), seule une onde longitudinale est présente à l'intérieur du tuyau, le champ rayonné est typiquement sphérique. Par contre à 4.5 kHz on observe clairement l'apparition d'un mode transverse qui affecte fortement le champ rayonné. Notons enfin le très bon accord entre les résultats mesurés et la simulation TLM.

Ce résultat peut être généralisé en considérant par exemple le coefficient de réflexion en pression :

$$R = \frac{1 - \frac{Z_r}{\rho c}}{1 + \frac{Z_r}{\rho c}} \quad \text{où} \quad Z_r = \frac{\iint_S P v_y^* dS}{\iint_S |v_y|^2 dS}$$

est l'impédance de rayonnement, v_y^* est le conjugué de la composante normale de la vitesse v_y (selon l'axe y), S est la surface rayonnante et ρ est la densité volumique. La Figure 8, qui présente les résultats mesurés ainsi que la simulation TLM, met en évidence la présence de deux fréquences de coupure correspondant à l'apparition du premier et du second mode transverses (4.3 kHz et 8.6 kHz). Pour ces fréquences on observe un soudain accroissement du coefficient de réflexion.

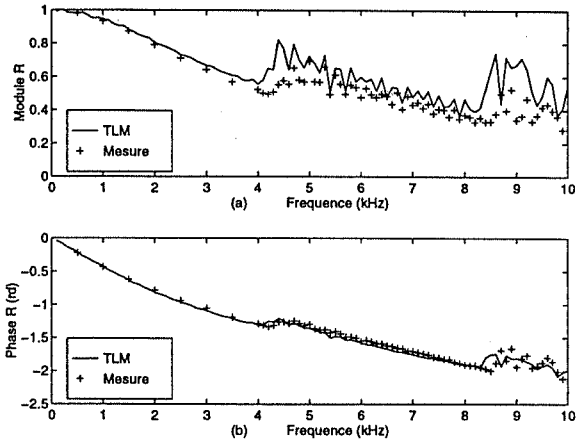


Figure 8 : Module et phase du coefficient du réflexion de la pression (— : simulation TLM, + : mesures).

4.2. Séparation des modes

Une analyse plus détaillée des résultats présentés dans la section 4.1. est difficile à réaliser du fait de la superposition des modes longitudinaux et transverses. Ceci s'illustre par les fortes fluctuations observées pour le module du coefficient de réflexion à partir de 4.3 kHz. Afin de pouvoir isoler le premier mode transverse nous avons réalisé deux simulations numériques. Dans la première, la source est située à l'extrémité fermée du guide sur l'axe du tuyau (x = 1.375 cm; y = 0.125 cm; z = 0.625 cm), dans la seconde la source est située dans un coin (x = 0.125 cm, y = 0.125 cm, z = 0.125 cm). D'un point de vue théorique on peut alors

montrer que, si dans le second cas le premier mode transverse est bien excité, dans le premier cas ce mode ne peut se propager (Pelorson et al., 1995). Ce résultat est confirmé par la simulation TLM (Figure 9).

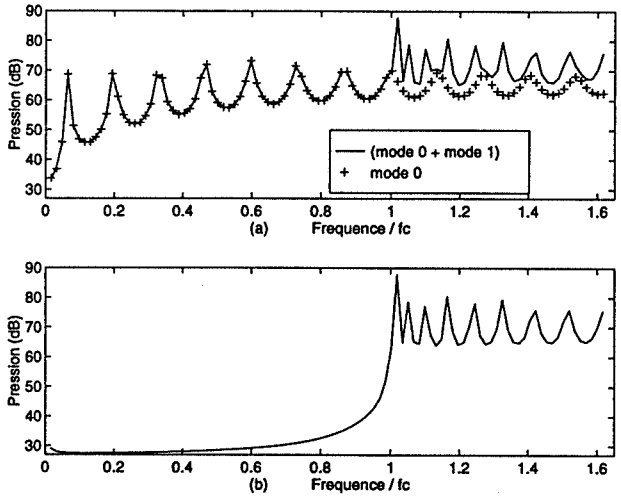


Figure 9 : Résultat de la simulation TLM: (a) module de la pression à l'extrémité du guide pour une source centrée (+) et en coin (—); (b) Mode transverse isolé par soustraction.
(fc est la fréquence de coupure du premier mode transverse).

La Figure 9(a) représente la pression calculée par simulation pour l'ensemble des modes présents (ici superposition du mode plan 0 avec le premier mode transverse 1) d'une part, et pour le mode plan uniquement d'autre part. Il est donc possible d'isoler la contribution du premier mode transverse en soustrayant la pression obtenue pour le mode plan à la pression obtenue pour les deux modes ensemble (voir Figure 9(b)).
Cette méthode permet donc aussi de calculer un coefficient de réflexion pour chaque mode. Le module et la phase des coefficients de réflexion ainsi obtenus sont présentés à la Figure 10.

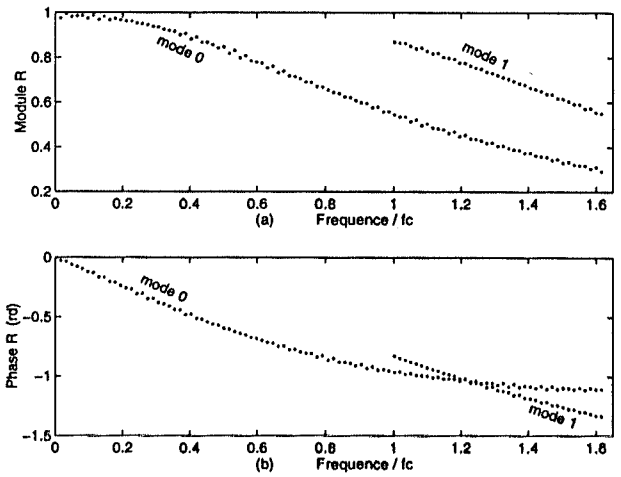


Figure 10: Module et phase du coefficient de réflexion modal. Le mode 0 correspond au mode plan, le mode 1 au premier mode transverse.

5. CONCLUSION

Dans cet article nous avons présenté une méthode de simulation numérique originale en acoustique. Cette méthode a été validée, soit par comparaison avec la théorie, soit par comparaison avec des résultats expérimentaux. De par sa précision et sa flexibilité, cette méthode s'avère un outil précieux pour l'analyse et la caractérisation de l'acoustique du conduit vocal. Ainsi nous avons montré, sur la base d'un modèle simple, l'importance des modes d'ordre supérieur non seulement en ce qui concerne la propagation acoustique dans le conduit mais aussi en ce qui concerne le rayonnement. Cet effet, déjà observé par d'autres auteurs (Matsuzaki et al., 1995, par exemple) n'avait pu jusqu'alors être quantifié de façon systématique. Nous avons également présenté une méthode qui permet d'isoler la contribution de chaque mode. Ces résultats et cette approche devraient à terme pouvoir être utilisés pour une synthèse de la parole de qualité.

6. REMERCIEMENTS

Cette étude a été partiellement financée par la fédération CNRS de laboratoires ELESA.

7. BIBLIOGRAPHIE

Bruneau M. (1983) *Introduction aux Théories de l'Acoustique,* Université du Maine, France.

Fant G. (1960) *Acoustic theory of speech production,* Mouton & Co., 's Gravenhage.

Hoefer W. J. R. (1989) *The Transmission Line Matrix (TLM) Method, Numerical technics for microwave and millimeter-wave passive structures,* Ed. Tatsuo Ltoh, Wiley Interscience.

Johns P.B. & Beurle R.L. (1971) Numerical solution of two dimensional scattering problems using a transmission line matrix, *Proc. IEEE,* 9, 118.

Lu C., Nakai T. & Suzuki H. (1993) Finite element simulation of sound transmission in vocal tract, *J. Acoust. Soc. Jpn,* 14, 63-72.

Matsuzaki H., Miki N. & Ogawa Y. (1995) 3-D FEM analysis of vocal tract models using elliptic tubes with volume radiation, *Proc. ICPhS,* 4, 440-443.

Pelorson X., Badin P., Motoki K., Miki N. & Plique M. (1995) On the radiation of sound at the lips during speech. Effects of lip geometry and of higher acoustical modes, *Proc. 15 th International Congress on Acoustics, Trondheim, Norway,* 4, 497-500.

Saguet P. (1985) Analyse des milieux guidés la méthode MTLM, *Thèse de Docteur d'État, INP - Grenoble.*

Saguet P. (1991) TLM Method for Three Dimensional Analysis of Microwave and mm-Structures, *International Workshop of German IEEE MTT/Ap* Chapter.

R1, R2 ET R3 : UN ENSEMBLE ROBUSTE DE PARAMÈTRES POUR LA CARACTÉRISATION DES ESPACES VOCALIQUES

Adrian NEAGU, Gérard BAILLY
Institut de la Communication Parlée - INPG & Université Stendhal
46 avenue Félix Viallet, 38031 Grenoble Cedex 1, France
E-Mail : neagu,bailly@icp.grenet.fr

ABSTRACT

This paper describes recognition and speaker normalisation experiments performed on vocalic realizations of 7 male and 6 female speakers. Two representations of vocalic targets are compared: formants vs resonances. Resonances achieve better performance : R1..2 set show only a slight performance reduction compared to R1..3 in recognitions tasks while it give the best representation for the vowel space in normalisation experiments.

Finally we test an algorithm for the automatic identification of R1..3 set based on R3 prediction from F1 value. It achieves 99.3% of correct identification of R3 on 959 vocalic samples.

Keywords : Formants affiliation, Vocalic space, Speaker normalisation.

1. INTRODUCTION

Les paramètres les plus utilisés pour caractériser la structure spectrale des voyelles orales sont les formants. Cet ensemble de paramètres ne facilite pas l'inversion acoustico-articulatoire car ils ne respectent pas les continuités spectrales (Badin, 1990). Un ré-étiquetage des formants selon leur affiliation majoritaire à une des cavités du conduit vocal a donc été proposé : on note R_{2*i} les formants associes à la cavité avant et R_{2*i-1} ceux de la cavité arrière. Les formants sont donc regroupés en deux familles qui présentent des mouvements d'ensemble cohérents (figure 1). Sur cette base, des représentations des espaces vocaliques qui tentent de "linéariser" la relation articulatori-acoustique ont été proposées (Bailly, 1991, 1993).

Dans la réalité, de nombreuses difficultés nous empêchent de déterminer avec précision les affiliations majoritaires des formants, notamment les problèmes liés au fort couplage pour les voyelles ouvertes. Nous utilisons ici une stratégie simplifiée de détermination des trois premières résonances :

- nous choisirons R1,R2,R3 parmi F1,F2,F3, même si, pour certaines réalisations de /i/, F4 est la première résonance de la cavité avant.
- R1 vaut toujours F1, même pour les voyelles qui risquent d'être focales en F1 et F2 (/a/, /u/, /o/).
- R2 = F2 et R3 = F3 sauf pour les voyelles antérieures.

Le plan R1,R2 empruntera les points du plan F1,F3 pour les voyelles /i/ et /e/ et les points du plan F1,F2 pour les autres voyelles. L'obtention des paramètres R1,R2,R3 se réduit donc à une détection de formants suivi de l'identification correcte de l'affiliation du F2, les affiliations des autres formants découlant des règles précédentes.

Nous avons réalisé au départ des tests supposant une identification idéale de R1,R2,R3 (l'expert a décidé qui est R3 connaissant les valeurs de F1,F2,F3 et la voyelle). Nous montrons ainsi que cet ensemble simplifié de paramètres apporte un gain sensible pour des tests de reconnaissance multilocuteur. Nous proposerons, bien sûr, par la suite une méthode

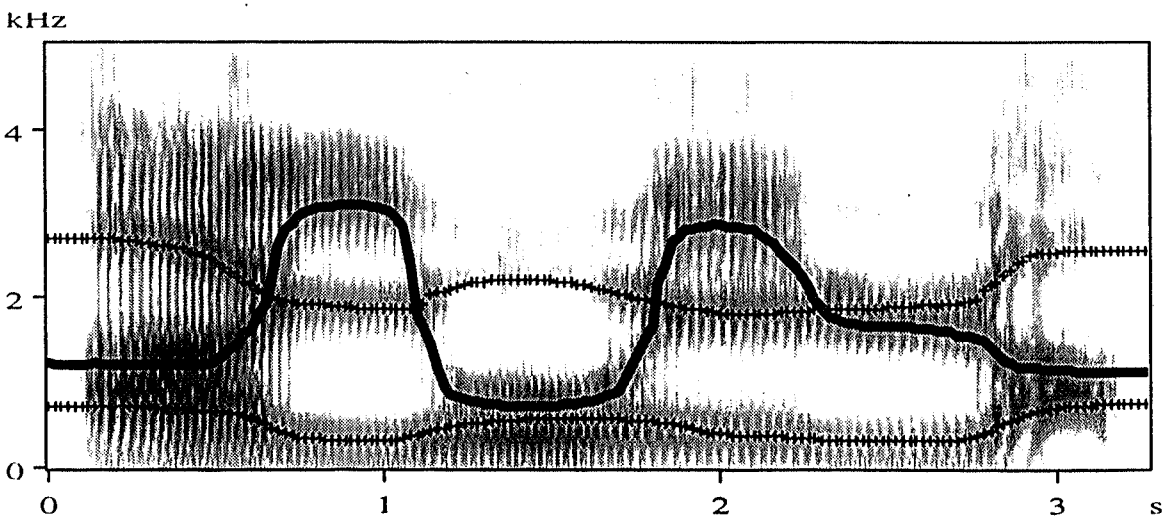


Fig.1: Exemple de suivi de résonances. La trajectoire du R2 est figurée en gras et superposée au sonagramme et à la famille R1,R3.

pour assurer le passage automatique de F1,F2,F3 vers R1,R2,R3, méthode basé sur la prédiction de la position de R3 selon la position de F1 et des caractéristiques spécifiques au locuteur.

Notons tout de suite que les taux de reconnaissance donnés dans cet article sont déterminés sur l'ensemble des réalisations y compris pour les tests de normalisation donc dans des contextes consonantiques variables[1] .

2. CORPUS DE TEST

Nous avons considéré une base largement utilisée de signaux de parole numériques (BDSONS). Nous avons sélectionné un sous-corpus présentant une bonne fréquence des toutes les voyelles orales (entre 30 pour les voyelles cardinales et 9 réalisations pour les voyelles centrales). Les deux corpus RAM01 et RAM10, constitués de mots monosyllabiques isolés, ont été analysés pour 6 locuteurs masculins de BDSONS. Nous avons donc au total 828 réalisations des différentes voyelles.

La banque de données comporte les valeurs des quatre premiers formants mesurés semi-automatiquement sur les cibles vocaliques. Un expert a précisé ces valeurs en regardant en même temps le sonagramme du signal et les valeurs proposées par deux méthodes automatiques (maxima du spectre FFT et racines du polynôme LPC)[2].

3. TEST DE RECONNAISSANCE

Les représentations formantiques et résonantiques ont été confrontées sur une tâche de décodage acoustico-phonétique. Nous avons choisi la classification bayesienne avec une fonction de probabilité gaussienne car son apprentissage nous fournit des représentations très lisibles (moyenne et matrice de covariance pour chaque classe) et facilement comparable aux données de la littérature. Nous présentons les résultats moyens (taux de reconnaissance, TR) d'un test croisé (apprentissage sur cinq locuteurs et test sur le sixième) pour divers jeux de paramètres (exprimés, tout d'abord, en Hertz) :

- Paramètres F1,F2 TR=84,4%
- Paramètres F1,F2,F3 TR=89,2%
- Paramètres R1,R2 TR=90,3%
- Paramètres R1,R2,R3 TR=91,0%

Les taux sont meilleurs pour les résonances en raison de l'introduction de connaissances a priori par notre méthode de détection des résonances (notamment toutes les confusions entre /e/ et /ɛ/ ont été supprimées). Nous remarquerons surtout la dégradation moindre pour le passage de trois à deux paramètres pour les résonances, preuve que la connaissance introduite est robuste. En effet, pour la moitié des locuteurs, TR obtenu à partir de R1..2 est meilleur que celui obtenu à partir de R1..3 tandis que les résultats pour F1..2 sont systématiquement inférieurs aux résultats pour F1..3.

Nous avons calculé aussi le TR sur les ensembles d'apprentissage (TA) afin de vérifier l'hypothèse de la distribution gaussienne et surtout afin de rendre compte de la capacité des différents systèmes de paramétrage à apprendre des prototypes vocaliques :

- Paramètres F1,F2 TA=90.75%
- Paramètres F1,F2,F3 TA=94,77%
- Paramètres R1,R2 TA=94,20%
- Paramètres R1,R2,R3 TA=95,67%

Nous avons repris le même test avec les données exprimées en barks. Nous avons calculé aussi pour tous les cas le nombre des erreurs graves (les erreurs où même le deuxième candidat proposé par le classifieur ne correspondait pas à la référence). Tous les indices ont montré les mêmes tendances : meilleurs résultats et dégradation plus faible au passage à deux paramètres pour les résonances.

Finalement, nous avons transposé les données de l'espace F1..3 vers l'espace CF1..3 par analyse discriminante canonique (respectivement de R1..3 vers CR1..3) :

- Paramètres CF1,CF2 TA=90,7%
- Paramètres CF1,CF2,CF3 TA=93,9%
- Paramètres CR1,CR2 TA=94,4%
- Paramètres CR1,CR2,CR3 TA=95,1%

Les résultats sont comparables mais la classification sur les composantes discriminantes présente l'inconvénient d'introduire des erreurs atypiques (confusions entre /e/ et /y/, par exemple).

Nous concluons que R1,R2 est un système de paramètres de dimension réduite qui garde l'essentiel de l'information phonétique de l'ensemble des voyelles orales. Notons toutefois que le troisième paramètre apporte toujours un gain informationnel, même s'il est réduit dans le cas des résonances.

4. MÉTHODE DE NORMALISATION

Les expériences précédentes ont été effectuées sans normalisation, les locuteurs ayant des espaces maximaux relativement proches. Dans cette seconde partie, la classification est précédée par une adaptation interlocuteur : cette transformation vise uniquement à recaler les espaces maximaux et ne nécessite donc aucune identification de phonèmes a priori. Trois homothéties sont déterminées pour F1..3 et trois autres pour R1..3.

[1] Certains travaux (Ferrari-Disner, 1980) utilisent par exemple la base Peterson-Barney qui ne donne que les valeurs formantiques de deux réalisations de plusieurs locuteurs dans un contexte fixe.

[2] Il faut souligner que certaines réalisations de /ɛ/ ont été effectivement prononcées fermées et auraient dû être étiquetées /e/ et vice-versa. Ces erreurs sont la cause d'un nombre non négligeable de mauvaises reconnaissances.

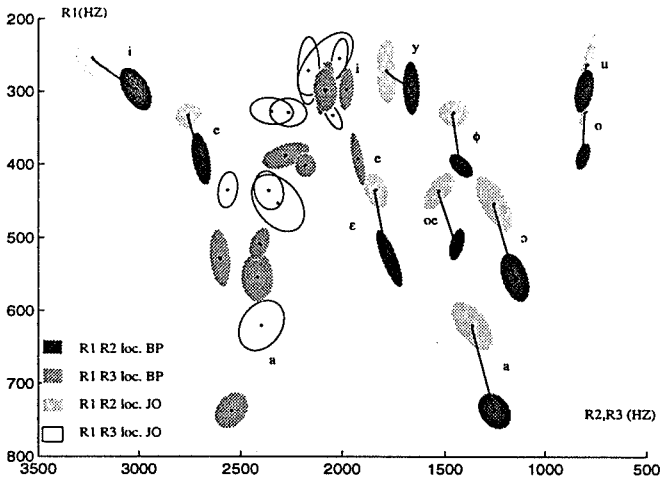


Fig.2: Positions et dispersions des voyelles pour les locuteurs BP et JO (projections dans les plans R1..2 et R1..3) avant normalisation. TR = 69,5% (TR = 63,8% pour F1..3).

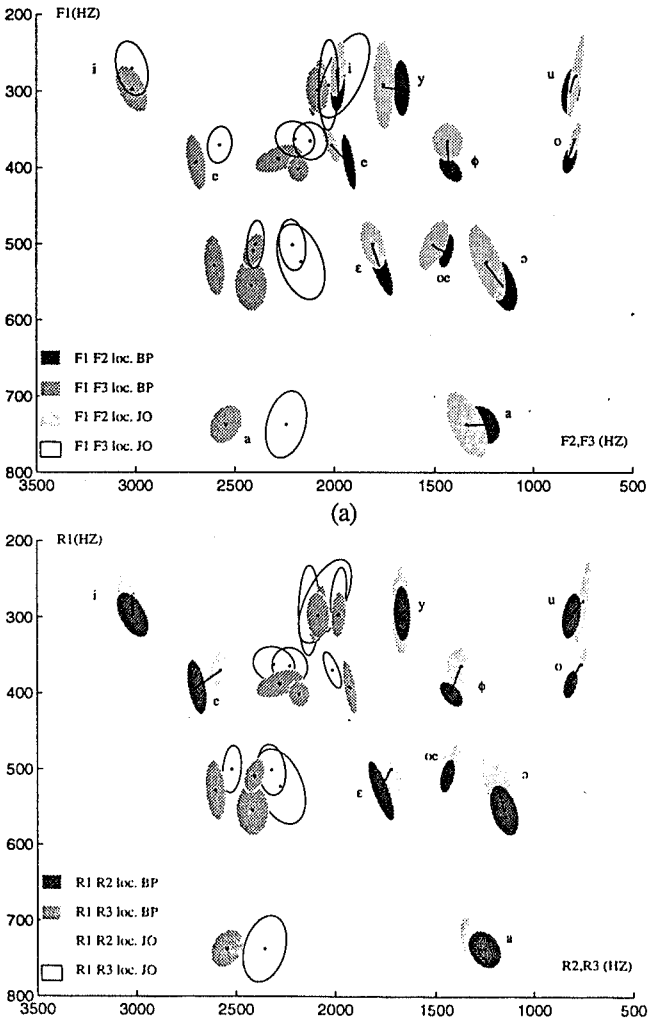


Fig.3.Positions et dispersions des voyelles pour les locuteurs BP et JO après normalisation (cible BP) sur (a) espace formantique maximal (TR = 78,2%) et (b) espace maximal des résonances (TR = 88,4%).

Elles ont toutes un centre commun pris comme la fréquence minimale observée (200Hz) et font correspondre les valeurs maximales[1] sur chaque composante pour les locuteurs source et cible. Cette méthode ne garantit pas la superposition exacte d'un nombre minimum de voyelles comme les méthodes classiques (Benedetto, 1992) (Ferrari-Disner, 1980) (Lobanov, 1971) : par exemple, les F2 et F3 du /a/ ne se superposent pas forcément.

5. TEST DE NORMALISATION

Pour le test, nous avons ajouté un septième locuteur masculin présentant un espace du R3 beaucoup plus haut en fréquence (F2(/i/) ≈ 2350 Hz) que les six autres (F2(/i/) ≈ 2000 Hz) et nous avons fait toute les combinaisons source - cible possibles. Nous obtenons TN, la moyenne de 6x7 = 42 taux de reconnaissance des voyelles du locuteur source avec un classifieur bayesien appris sur les voyelles du locuteur cible. Nous avons fait, tout d'abord, un test sans normalisation aucune comme référence :

- Paramètres F1,F2 TN=69,1%
- Paramètres F1,F2,F3 TN=70,7%
- Paramètres R1,R2 TN=77,4%
- Paramètres R1,R2,R3 TN=74,7%

On identifie ainsi très vite les locuteurs atypiques (TR moyen pour cible JS = 37% en F1..3) et les locuteurs qui prononcent avec peu de précision (TR moyen pour source JO = 58,3% en F1..3). Remarquons que le plan R1,R2 est déjà le meilleur pour une simple superposition multilocuteur.

Le même test, mais après rapprochement du locuteur source du locuteur cible par la méthode décrite plus haut, donne :

- Paramètres F1,F2 TN=78,2%
- Paramètres F1,F2,F3 TN=76,2%
- Paramètres R1,R2 TN=84,3%
- Paramètres R1,R2,R3 TN=78,3%

A la différence du test de reconnaissance, nous observons que le troisième paramètre peut aussi introduire du bruit. Il s'agit surtout de F3 (ou bien R3) des voyelles demi - ouvertes qui ne conserve même pas la même topologie d'un locuteur à l'autre. L'augmentation de TN après normalisation a été obtenue en réduisant la dynamique des résultats (les très bons taux de reconnaissance ont diminué un peu mais, surtout, les plus mauvais taux se sont beaucoup améliorés).

Comme pour le test de reconnaissance sans normalisation, le changement d'échelle (Barks) montre la même relation entre les différents systèmes de paramètres.

Finalement, nous avons ajouté 6 locuteurs féminins à notre base et nous avons fait les tests de

[1] Nous avons pris comme valeur maximale de F1 la moyenne sur F1 de la cible du /a/, pour F2 celle sur F2 du /i/ et pour F3 celle sur F3 du /i/. Finalement, pour les résonances, ce sont les mêmes homothéties que pour les formants mais elles agissent différemment sur les données.

reconnaissance faisant intervenir une cible et une source de sexe différent. Nous obtenons en utilisant le meilleur système de paramètres (R1, R2) TN = 55,7% avant et TN = 84,9 % après normalisation. Ce résultat montre que notre procédure de normalisation peut assurer en décodage acoustico-phonétique le même niveau de performance quelque soit le groupe de locuteurs considéré.

6. PRÉDICTION DU R3

Tous les points du plan R1,R3 de six locuteurs masculins de BDSONS semblent s'encadrer dans une bande de 600 Hz autour de la droite de régression (R3 = 1,08 * R1 + 1818 Hz). Les voyelles les plus écartées de cette droite sont /e/ et /o/, deux voyelles très proches en R1 mais très éloignées en R3. Pour palier le manque de corrélation entre R1 et R3 dans cette situation, nous avons choisi de remplacer la droite de régression par deux segments de droite. A cet effet, nous avons choisi pour chaque locuteur un prédicteur R3p de la valeur de R3 selon la stratégie suivante :

Si F1 < (F1moyen(/ø/) + F1moyen(/œ/))/2 alors R3p = moyenne(F2(/i/),F3(/y/),F3(/u/)) sinon R3p = F3moyen(/a/).

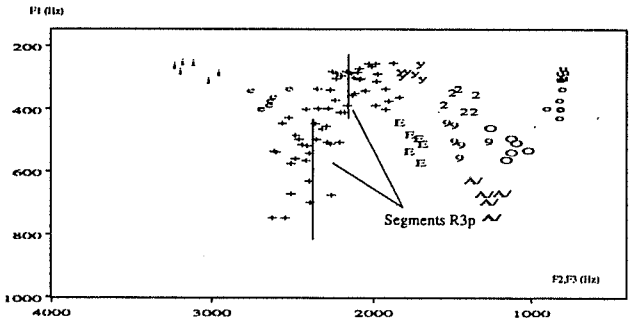


Fig. 4 Position des voyelles pour les six locuteurs dans les plans F1 F2 et F1 F3 superposés. Les points identifiés comme appartenant à la région R1,R3 ont été représentés par des croix et les autres par leur étiquettes phonétiques (en code SAM-BDSONS).

L'algorithme d'identification des résonances considère R3 comme le formant le plus proche de R3p. L'algorithme atteint 99.3% de réussite sur les 7 locuteurs masculins (959 voyelles) et 98,9% sur les 6 locuteurs féminins (827 voyelles). Toutes les erreurs sont attribuées à l'imprécision d'étiquetage phonétique sur /e/ versus /ɛ/.

Ce modèle est robuste : si R3p est calculé sur tous les locuteurs masculins confondus (R3p = 2150 Hz si F1<420 Hz sinon R3p = 2750 Hz), le taux de réussite est encore de 97,9%.

Il est important à noter que la détermination de R3 est sensible à la stratégie d'opposition d'ouverture propre au locuteur et non-prédictible à partir de son espace maximal.

7. CONCLUSIONS ET PERSPECTIVES

Nous observons sur nos données que la séparation de résonances de la cavité avant (R2) peut améliorer l'identification des voyelles dans un système de D.A.P. (performances comparables avec l'analyse discriminante canonique mais avec un calcul moins coûteux) et peut servir comme une bonne base pour une normalisation des locuteurs en vue de la reconnaissance.

Ces résultats vont nous permettre d'introduire une forte connaissance a priori sur la structure formantique et donc de contraindre un suivi automatique de formants.

Il est à noter que cette représentation remet aussi en question la représentation des consonnes. Notamment, les équations de locus pourront prendre en compte des éventuels changements d'affiliations tels que ceux présentés dans (Eek, 1995).

RÉFÉRENCES

Badin, P., Boë, L.J., Perrier, P., and Abry, C. (1990) Acoustic considerations upon formant convergence. *Journal of the Acoustical Society of America,* 87(3), 1290-1300

Bailly, G. and Guerti, M. (1991) Synthesis-by-rule for french. In *12th International Congress of Phonetic Sciences,* pages 506-509

Bailly, G. (1993) Resonances as possible representation of speech in the auditory-to-articulatory transform. *EUROSPEECH*, pages 1511-1514

Bailly, G. (1995) Caracterisation of formant trajectories by tracking vocal tract resonances. In Beekmans, R., Jospa, P., Schoentgen, J., and Serniclaes, W.,editors, *Levels in speech communication : relations and interactions*, pages 91-102 Elsevier, Amsterdam

Benedetto, M.G.D. and Liénard, J.S. (1992) Extrinsic normalization of vowel formant values based on cardinal vowels mapping. In *International Congress on Speech and Language Processing*, pages 579-582, Banff - Alberta, Canada

Eek A. and Meister E. (1995) The perception of stop consonants : locus equations and spectral integration, In *International Congress of Phonetic Sciences*, volume 1, pages 18-21, Stockholm, Sweden

Ferrari-Disner, S. (1980) Evaluation of vowel normalization procedures. *Journal of the Acoustical Society of America*, 67, 253-261

Lobanov, B. (1971) Classification of russian vowels spoken by different speakers. *Journal of the Acoustical Society of America*, 49, 606-608

Peterson, G. and Barney, H. (1952) Control methods used in a study of vowels. *Journal of the Acoustical Society of America,* 20, 528-535

SOUS-ESPACES DE PROJECTION DE SEQUENCES DE TRAMES ACOUSTIQUES POUR L'ANALYSE ET LA RECONNAISSANCE DE PAROLE

Frédéric BIMBOT [2][1], Enrico BOCCHIERI [1], Bishnu ATAL [1]

(1) AT&T - Bell Laboratories - Speech Research Dept, 600 Mountain Ave., Murray Hill, NJ 07974, Etats-Unis.
(2) ENST - Dépt Signal, CNRS - URA 820, 46 Rue Barrault, 75634 Paris cedex 13, France, Union Européenne.

ABSTRACT

The use of dynamic (delta and delta-delta) cepstral vectors or of RASTA parameters has been shown to be a significant factor of improvement for speech recognition. These approaches can be gathered under a common formalism and understood as the projection of acoustic frame sequences onto specific subspaces. We propose here a third alternative, where the projection subspace is derived from the acoustic data themselves, using principal component analysis. The coefficients obtained from these *principal temporal components* provide slightly better results than conventional and more arbitrary dynamic parameters, for a reference speech recognition task.

1. INTRODUCTION

Une pratique courante en analyse de parole pour la reconnaissance consiste à adjoindre aux vecteurs de paramètres acoustiques des vecteurs auxiliaires, généralement présentés comme des approximations de la dérivée et de la dérivée seconde de ces mêmes paramètres. Ainsi, on complète les coefficients cepstraux par des coefficients *delta*-cepstraux et *delta-delta*-cepstraux.

Expérimentalement, les taux de reconnaissance obtenus par ces paramètres augmentés sont bien meilleurs que ceux observés avec les vecteurs cepstraux seuls. On interprète souvent ce résultat comme une indication de l'importance de l'information acoustique dynamique, les delta- et les delta-delta-paramètres pouvant être compris comme caractéristiques de la vitesse et de l'accélération de l'évolution des paramètres acoustiques.

Dans cet article, nous exposons deux interprétations supplémentaires de cette approche : l'une en termes de filtrage, l'autre en termes de projection. La première reformule le calcul des vecteurs auxiliaires comme le résultat de produits de convolution et permet de relier les delta-paramètres aux paramètres RASTA. La seconde suggère de dériver le calcul des vecteurs augmentés directement à partir des données acoustiques d'apprentissage en ne conservant que l'information statistiquement essentielle dans une séquences de trames acoustiques de longueur fixe.

On décrit alors un procédé de construction des paramètres augmentés par projection de séquences de vecteurs acoustiques sur les premiers vecteurs propres de la matrice d'autocorrélation cumulée des composantes de ces vecteurs, calculée à long terme. Ces vecteurs propres (*composantes temporelles principales* du spectre) présentent une forte parenté avec les fonctions convoluantes utilisées pour calculer les delta- et les delta-delta-paramètres. Cependant, elles peuvent être mieux approximées par des fonctions trigonométriques.

En substituant les nouveaux paramètres ainsi obtenus à ceux utilisés dans un système de reconnaissance de parole de référence à base de paramètres cepstraux, delta-cepstraux et delta-delta-cepstraux, on observe une très légère amélioration des performances.

2. FORMULATION

Soit $\mathbf{X}_1^n = [\mathbf{x}_t]_{1 \leq t \leq n}$ une suite de n paramètres acoustiques de dimension m. On notera x_{kt} la $k^{\text{ième}}$ coordonnée du vecteur $\mathbf{x}_t$ (avec $1 \leq k \leq m$).

Si l'on note maintenant $\dot{\mathbf{x}}_t$ et $\ddot{\mathbf{x}}_t$ des approximations respectives de la dérivée et de la dérivée seconde des paramètres acoustiques à l'instant t, l'adjonction des delta- et des delta-delta-paramètres revient à former une matrice $\mathbf{Y}_t$ obtenue par juxtaposition des vecteurs $\mathbf{x}_t$, $\dot{\mathbf{x}}_t$ et $\ddot{\mathbf{x}}_t$:

$$\mathbf{Y}_t = \left[\mathbf{x}_t \quad \dot{\mathbf{x}}_t \quad \ddot{\mathbf{x}}_t \right] \tag{1}$$

Cette matrice est ici de dimension $m \times 3$. En pratique, on applique des facteurs d'échelle (indépendants du temps) à chaque vecteur, dont les valeurs sont ajustées plus ou moins empiriquement pour optimiser les performances du système de reconnaissance.

Plus généralement, on notera :

$$\mathbf{Y}_t = \left[\mathbf{y}_t^{(0)} \quad \mathbf{y}_t^{(1)} \quad \mathbf{y}_t^{(2)} \quad \ldots \quad \mathbf{y}_t^{(p-1)} \right] \tag{2}$$

la matrice formée par juxtaposition des vecteurs $\mathbf{y}_t^{(i)}$ (avec $0 \leq i \leq p-1$), chacun de ces vecteurs de dimension m étant obtenu en appliquant une fonction F_i à la séquence $\mathbf{X}_{t-q}^{t+q} = [\mathbf{x}_{t-q} \ldots \mathbf{x}_{t+q}]$, c'est-à-dire au vecteur $\mathbf{x}_t$ et à ses contextes droits et gauches de longueur q de part et d'autre. La matrice $\mathbf{Y}_t$ est donc de dimension $m \times p$ et est calculée sur $2q+1$ vecteurs consécutifs[1].

Dans le reste de cet article, on suppose que la fonction F_i peut se mettre sous la forme :

$$\mathbf{y}_t^{(i)} = F_i(\mathbf{X}_{t-q}^{t+q}) = \mathbf{X}_{t-q}^{t+q} \cdot \mathbf{h}_i^T \tag{3}$$

où $\mathbf{h}_i$ est un vecteur ligne de $2q+1$ valeurs réelles et de norme unité. On notera $h_{i\tau}$ la coordonnée courante de $\mathbf{h}_i$ et on convient que $-q \leq \tau \leq +q$.

En notation matricielle, on peut écrire :

$$\mathbf{Y}_t = \mathbf{X}_{t-q}^{t+q} \cdot \mathbf{H}^T \tag{4}$$

où $\mathbf{H}^T = [\mathbf{h}_0^T \, \mathbf{h}_1^T \ldots \mathbf{h}_{p-1}^T]$ est une matrice $(2q+1) \times p$ dont les colonnes sont normées.

Finalement, l'application de facteurs d'échelle aux vecteurs $\mathbf{y}_t^{(i)}$ revient à calculer :

$$\mathbf{Z}_t = \mathbf{Y}_t \cdot \Lambda \tag{5}$$

où Λ est une matrice diagonale $p \times p$ dont l'élément courant sera noté λ_i. Ce sont les coefficients de $\mathbf{Z}_t$ qui sont présentés en entrée du système de reconnaissance.

[1] On peut naturellement généraliser cette notation aux séquences de longueur paire et aux contextes asymétriques.

3. INTERPRETATIONS

• cinématique

En pratique, les vecteurs $\dot{\mathbf{x}}_t$ et $\ddot{\mathbf{x}}_t$ de l'équation (1) sont approximés par des combinaisons linéaires des vecteurs $\mathbf{x}_{t-q}$ à $\mathbf{x}_{t+q}$, le contexte $[t-q,\, t+q]$ couvrant en général un intervalle de temps de l'ordre de 50 ms.

Par exemple, dans le système de référence décrit au paragraphe 5, on calcule, à partir de trames centisecondes :

$$\dot{\mathbf{x}}_t = \frac{1}{10}\sum_{\tau=-2}^{\tau=2} \tau \,.\, \mathbf{x}_{t-\tau} \qquad \ddot{\mathbf{x}}_t = \frac{1}{2}\sum_{\tau=-1}^{\tau=1} \tau \,.\, \dot{\mathbf{x}}_{t-\tau} \tag{6}$$

Ici, la dérivée $\dot{\mathbf{x}}_t$ est approximée par la corrélation entre 5 trames consécutives et une rampe linéaire de longueur 5, et la dérivée seconde $\ddot{\mathbf{x}}_t$ s'obtient en redérivant $\dot{\mathbf{x}}_t$ en utilisant une rampe de longueur 3.

En posant :

$$\begin{array}{rcl} \eta_0 & = & w_0 \,[\; 0 \quad 0 \quad 0 \quad 1 \quad 0 \quad 0 \quad 0 \;] \\ \eta_1 & = & w_1 \,[\; 0 \quad -2 \quad -1 \quad 0 \quad 1 \quad 2 \quad 0 \;] \\ \eta_2 & = & w_2 \,[\; 2 \quad 1 \quad -2 \quad -2 \quad -2 \quad 1 \quad 2 \;] \end{array} \tag{7}$$

$$\text{avec} \quad w_0 = 1 \quad w_1 = 1/\sqrt{10} \quad w_2 = 1/\sqrt{22}$$

la façon de calculer $\mathbf{Y}_t = [\mathbf{x}_t \;\; \dot{\mathbf{x}}_t \;\; \ddot{\mathbf{x}}_t]$ apparaît comme un cas particulier (aux facteurs d'échelle près[2]) de l'équation (3) ($p = 3$, $q = 3$, $\mathbf{h}_i = \eta_i$). Les valeurs numériques et la représentation graphique des η_i sont données sur la figure 1.

On peut également remplacer $\mathbf{x}_t$, $\dot{\mathbf{x}}_t$ et $\ddot{\mathbf{x}}_t$ par des valeurs obtenues après approximation polynomiale de la trajectoire locale, par exemple en corrélant la séquence $\mathbf{X}_{t-q}^{t+q}$ avec les vecteurs α_i vérifiant :

$$\begin{aligned} \alpha_{0\tau} &= u_0 \qquad \alpha_{1\tau} = u_1\,\tau \\ \alpha_{2\tau} &= u_2\,[\tau^2 - \frac{1}{3}q(q+1)] \end{aligned} \tag{8}$$

En choisissant :

$$u_0 = \sqrt{\frac{1}{2q+1}} \quad u_1 = \sqrt{\frac{12}{(2q+2)(2q+1)(2q)}}$$

$$u_2 = \sqrt{\frac{180}{(2q+3)(2q+2)(2q+1)(2q)(2q-1)}}$$

les α_i (représentés sur la figure 1) forment un système de vecteurs orthonormés, ce qui n'est pas le cas des η_i, seulement normés.

• filtrage

En réécrivant l'équation (3) sous la forme :

$$y_{kt}^{(i)} = \sum_{\tau=-q}^{\tau=q} h_{i\tau} \,.\, x_{k(t+\tau)} \tag{9}$$

la $k^{\text{ième}}$ coordonnée du vecteur $\mathbf{y}_t^{(i)}$ apparaît comme le résultat du filtrage du signal scalaire x_{kt} par un filtre de réponse impulsionnelle $h_{i(-\tau)}$.

Cette interprétation amène à s'intéresser à la fonction de transfert des filtres correspondants. A titre d'illustration, on a représenté sur la figure 1 la magnitude signée de la Transformée de Fourier des réponses impulsionnelles $\eta_{i\tau}$ et $\alpha_{i\tau}$ retournées. Cette magnitude est obtenue à partir d'une Transformée de Fourier Rapide sur 256 points, dont on ne représente que les 129 premières valeurs de la partie réelle (pour les vecteurs symétriques) ou de la partie imaginaire (pour les vecteurs antisymétriques). La figure 1 indique également la fréquence centrale des différents filtres, sur la base d'une analyse spectrale au rythme de 100 trames par secondes.

Ainsi, le calcul des delta-paramètres sur 5 trames centisecondes consécutives s'apparente à un filtrage temporel passe-bande peu sélectif des paramètres spectraux autour de 14 Hz environ[3]. Ce filtrage s'accompagne d'un déphasage de $\pi/2$.

L'interprétation en termes de filtrage met en évidence le lien étroit entre les paramètres delta et les paramètres RASTA [1]. Ces derniers sont obtenus par filtrage passe-bande des séquences de trames acoustiques et apportent, lorsque les caractéristiques des filtres sont bien choisies, une amélioration et une robustesse accrue de la reconnaissance. Le formalisme de l'équation (3) englobe donc l'approche RASTA, pour des filtres à réponse impulsionnelle finie.

• géométrique

L'équation (3) peut également s'interpréter en termes de projection. Ainsi, l'élément $y_{kt}^{(i)}$ du vecteur $\mathbf{y}_t^{(i)}$ apparaît dans l'équation (9) comme le produit scalaire du vecteur ligne $\mathbf{x}_{k\,[t-q\;t+q]} = [x_{k(t-q)} \cdots x_{k(t+q)}]$ avec le vecteur $\mathbf{h}_i$. Si le système de vecteurs $\mathbf{h}_i$ est orthonormé, la projection $\hat{\mathbf{x}}_{k\,[t-q\;t+q]}$ de $\mathbf{x}_{k\,[t-q\;t+q]}$ sur le sous-espace engendré par les vecteurs $\mathbf{h}_i$ s'écrit :

$$\hat{\mathbf{x}}_{k\,[t-q\;t+q]} = \sum_{i=0}^{p-1} y_{kt}^{(i)} \,.\, \mathbf{h}_i \tag{10}$$

soit, en forme matricielle, et en utilisant l'équation (4) :

$$\hat{\mathbf{X}}_{t-q}^{t+q} = \mathbf{Y}_t \,.\, \mathbf{H} = \mathbf{X}_{t-q}^{t+q} \,.\, \mathbf{H}^T \,.\, \mathbf{H} \tag{11}$$

Cette interprétation incite à rechercher un système de vecteurs $\mathbf{h}_i$ à partir des données $\mathbf{X}_1^n$, de telle sorte que le sous-espace associé soit celui dans lequel la perte d'information par projection est globalement minimale au cours du temps.

4. PROJECTION DÉDUITE DES DONNÉES

• principe

Soit $\mathbf{X}$ une matrice de dimension $m \times (2q+1)$ et $\mathbf{H}$ une matrice de projection orthogonale sur un sous-espace de dimension p. Les p lignes de la matrice $\mathbf{H}$ qui minimise l'erreur quadratique entre les vecteurs de $\mathbf{X}$ et les vecteurs projetés de $\hat{\mathbf{X}} = \mathbf{X}\mathbf{H}^T\mathbf{H}$ s'obtiennent comme les p vecteurs propres associés aux p plus grandes valeurs propres de la matrice (symétrique) $\mathbf{S} = \mathbf{X}^T\mathbf{X}$.

En pratique, on décompose $\mathbf{S}$ sous la forme :

$$\mathbf{S} = \mathbf{G}\,\mathbf{D}\,\mathbf{G}^{-1} \tag{12}$$

où $\mathbf{D}$ est une matrice diagonale dont les éléments sont rangés par ordre décroissant, et où $\mathbf{G}$ est une matrice orthonormale ($\mathbf{G}^{-1} = \mathbf{G}^T$). La matrice $\mathbf{H}^T$ s'obtient alors comme les p premières colonnes de $\mathbf{G}$.

On généralise cette approche en définissant :

$$\bar{\mathbf{S}} = \frac{1}{n-2q}\sum_{t=q+1}^{t=n-q} \left(\mathbf{X}_{t-q}^{t+q}\right)^T \mathbf{X}_{t-q}^{t+q} \tag{13}$$

et en choisissant comme sous-espace de projection celui engendré par les p ($\leq q$) premiers vecteurs propres de $\bar{\mathbf{S}}$. Ces *composantes temporelles principales* seront notées γ_i dans la suite de l'article.

[2] Les valeurs des λ_i sont données dans le paragraphe 5.

[3] Cette valeur est du même ordre que le débit phonétique.

On montre que l'élément courant $s_{\tau\theta}$ $(-q \leq \tau, \theta \leq q)$ de la matrice $\bar{\mathbf{S}}$ s'approxime, pour $n \gg q$, sous la forme :

$$s_{\tau\theta} \approx \sum_{k=1}^{k=m} \left(\frac{1}{n} \sum_{t=1+|\theta-\tau|}^{n} x_{kt}\, x_{k(t-|\theta-\tau|)} \right) \qquad (14)$$

La matrice $\bar{\mathbf{S}}$ est donc, à peu de choses près, la somme des m matrices d'autocorrélation d'ordre $2q+1$ calculées pour chaque composante k de $\mathbf{X}_1^n$ et a notamment une structure de Toeplitz. Les vecteurs γ_i sont donc peu différents des p premiers vecteurs propres de cette matrice d'autocorrélation cumulée.

• **application**

La figure 1 donne une représentation de l'allure temporelle et du spectre des 3 premiers vecteurs γ_i obtenus pour $q = 2$ selon le principe ci-dessus à partir d'approximativement 6 minutes de parole.

Pour cette expérience, nous avons utilisé un sous-ensemble de l'ensemble d'apprentissage de la base de données ATIS [2]. Il s'agit de 66 phrases (soit 378.26 secondes de parole) échantillonnées à 16 kHz sur 16 bits, produites par 22 locuteurs, masculins et féminins, s'exprimant en anglais américain. Les vecteurs $\mathbf{x}_t$ sont constitués de 12 coefficients MFCC, calculés sur des trames de 20 ms à raison de 100 trames par seconde[4], avec une préaccentuation de 0.99 et un fenêtrage de Hamming.

On observe une relative ressemblance entre les composantes temporelles principales γ_i et les vecteurs polynomiaux α_i définis au paragraphe 3, ainsi qu'entre leurs spectres. Cependant, les γ_i sont mieux modélisés par des fonctions trigonométriques, que nous décrivons maintenant..

• **modélisation**

Soient, pour $0 \leq i \leq 2$, les vecteurs :

$$\beta_{i\tau} = v_i \cos\left[i \left(\frac{\tau}{2q+1} - \frac{1}{2} \right) \pi \right] \qquad (15)$$

$$\text{avec} \qquad v_0 = \frac{1}{\sqrt{2q+1}} \qquad v_1 = v_2 = \frac{\sqrt{2}}{\sqrt{2q+1}}$$

les vecteurs β_i (représentés sur la figure 1 pour $q = 2$) forment un système de vecteurs orthonormés. Ils fournissent d'excellentes approximations des vecteurs γ_i tant dans le domaine temporel que dans le domaine spectral.

Plus précisément, la différence absolue $| \gamma_{i\tau} - \beta_{i\tau} |$ entre les vecteurs γ_i obtenus expérimentalement pour $1 \leq q \leq 5$ et les modèles proposés β_i n'a jamais excédé 4.10^{-2}.

5. EXPÉRIENCES

La validité des composantes temporelles principales γ_i a été testée en les substituant aux α_i au sein d'un système de référence de reconnaissance de parole, et en comparant les performances obtenues sur la base de données ATIS.

Ce système de référence est un système à base de modèles de Markov cachés à densités de probabilité continues, composé de 47 modèles de phonèmes[5] indépendants du contexte (et du sexe). Chaque phonème est modélisé par 3 états et chaque état par 32 distributions gaussiennes. Le système de reconnaissance utilise un modèle de langage bigramme.

L'ensemble d'apprentissage contient 21 000 phrases spontanées (soit 208 000 mots) couvrant un vocabulaire de 1 530 mots. L'ensemble de test est composé de 998 phrases (10 081 mots) prononcées par des locuteurs différents de ceux de l'apprentissage, et dont 99.8 % des occurences de mots font partie du vocabulaire d'apprentissage.

Les paramètres acoustiques sont 12 coefficients MFCC obtenus dans les conditions d'analyse décrites dans le paragraphe 4, ainsi qu'un coefficient d'énergie en dB.

Pour l'expérience I, on adjoint à ces 13 coefficients les coefficients delta et delta-delta obtenus à partir des vecteurs α_1 et α_2 dont l'expression est donnée au paragraphe 3, avec des facteurs d'échelle $\lambda_1 = 1.18$ et $\lambda_2 = 0.66$ ($\lambda_0 = 1.00$). Pour l'expérience II, on remplace ces 39 coefficients par 39 autres obtenus à partir des γ_i ($q = 2$, $p = 3$). On conserve les mêmes facteurs d'échelle que pour l'expérience I.

La table 1 donne les performances obtenues pour les 2 expériences en termes de taux d'erreur (de précision) sur les mots, ce taux étant calculé comme le nombre de substitutions, insertions et omissions rapporté au nombre de mots de test.

	Expérience I vecteurs α_i	Expérience II vecteurs γ_i
Taux d'erreur	7.6 %	7.4 %

Table 1: Taux d'erreur sur ATIS pour des sous-espaces de projection définis empiriquement (α) et à partir de données d'apprentissage (γ) .

L'expérience II donne des performances légèrement meilleures que l'expérience I[6], mais l'écart n'est peut-être pas significatif. Ce résultat montre néanmoins que le procédé de construction des vecteurs convoluants à partir des données permet d'aboutir à une base de projection au moins aussi satisfaisante que celle définie empiriquement[7].

6. CONCLUSIONS ET PERSPECTIVES

Le travail présenté dans cet article ouvre une nouvelle piste dans la recherche de procédures non-empiriques pour le choix de paramètres d'analyse de la parole. Le formalisme proposé permet d'établir un lien entre l'interprétation cinématique des paramètres delta, les fonctions de convolution des paramètres RASTA et des propriétés statistiques intrinsèques aux données.

Le procédé proposé pour générer les vecteurs de projection est simple à mettre en œuvre et les résultats expérimentaux confirment sa validité sur la tâche testée. Des expériences sur d'autres tâches, d'autres langues et pour d'autres applications viendront confirmer ou infirmer le bien-fondé de l'approche.

7. REFERENCES

[1] H. Hermansky, N. Morgan, A. Bayya and P. Kohn (1991). *Compensation for the effect of the communication channel in auditory-like analysis of speech (RASTA-PLP).* Eurospeech 91, pp. 1367-1370.

[2] MADCOW (1992). *Multidata collection for a spoken language corpus.* DARPA Workshop on Speech and Natural Language. Harriman, 1992, pp. 7-14.

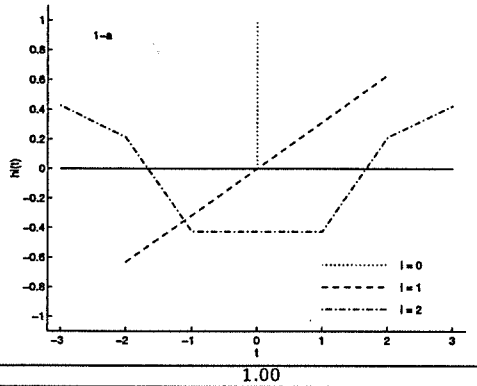
[3] E. Bocchieri, G. Riccardi and J. Anantharaman (1995). *The 1994 AT&T ATIS Recognizer.* ARPA Spoken Language Systems Technology Workshop, Austin, 1995, pp. 265-269.

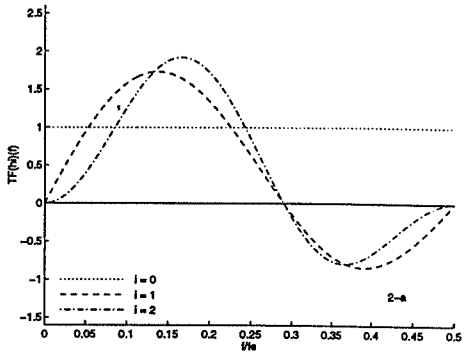
[4] On a donc $n = 37826$.

[5] Et 3 modèles de silence / bruit.

[6] Le score de l'expérience I est néanmoins inférieur à celui que l'on obtient avec un système de reconnaissance plus élaboré : le meilleur système développé à Bell Labs permet de descendre à 3.5 % d'erreur sur la même tâche [3].

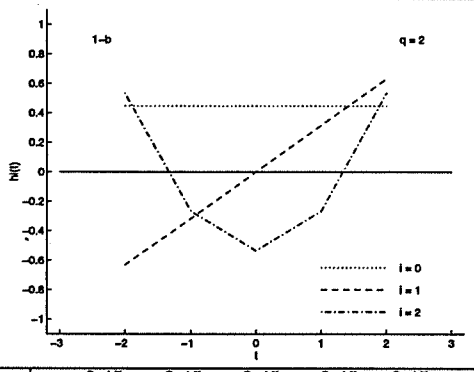
[7] Néanmoins, le choix de p, q et l'ajustement des facteurs d'échelle demeurent empiriques.

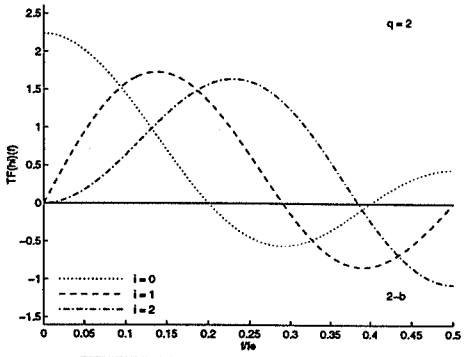




η_0				1.00			
η_1		-0.63	-0.32	0.00	0.32	0.63	
η_2	0.43	0.21	-0.43	-0.43	-0.43	0.21	0.43

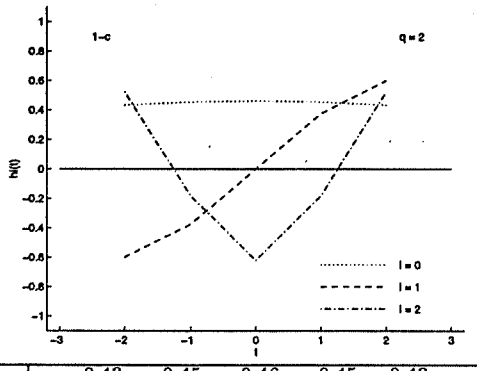
Fréquence du lobe principal		
$i = 0$	$i = 1$	$i = 2$
—	13.7 Hz	16.8 Hz

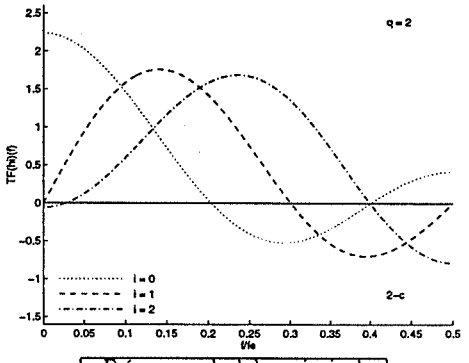




α_0	0.45	0.45	0.45	0.45	0.45
α_1	-0.63	-0.32	0.00	0.32	0.63
α_2	0.53	-0.27	-0.53	-0.27	0.53

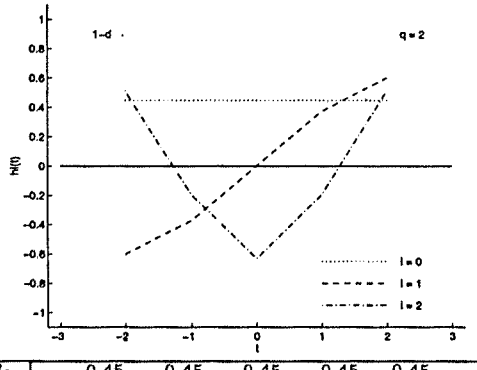
Fréquence du lobe principal		
$i = 0$	$i = 1$	$i = 2$
0 Hz	13.7 Hz	23.0 Hz

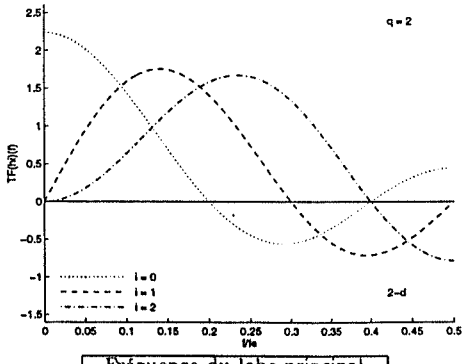




γ_0	0.43	0.45	0.46	0.45	0.43
γ_1	-0.60	-0.38	0.00	0.38	0.60
γ_2	0.52	-0.18	-0.62	-0.18	0.52

Fréquence du lobe principal		
$i = 0$	$i = 1$	$i = 2$
0 Hz	14.1 Hz	23.4 Hz





β_0	0.45	0.45	0.45	0.45	0.45
β_1	-0.60	-0.37	0.00	0.37	0.60
β_2	0.51	-0.20	-0.63	-0.20	0.51

Fréquence du lobe principal		
$i = 0$	$i = 1$	$i = 2$
0 Hz	14.1 Hz	23.4 Hz

Figure 1: A gauche : représentation graphique et valeur approchée des vecteurs η, α, γ et β ($p = 3$, $q = 2$ sauf pour η). A droite : fonctions de transfert correspondantes et fréquence du maximum du spectre (sur la base d'une analyse centiseconde).

UNE EVALUATION EXPERIMENTALE DES PERFORMANCES DE PLUSIEURS PDA's EN PRESENCE DE SEPT NIVEAUX D'UN BRUIT GAUSSIEN

Malika BOUDRAA*, Bachir BOUDRAA* , Bernard GUERIN**

*Institut d'électronique, USTHB , BP 32 EL-ALIA , ALGER , ALGERIE

**I.C.P./ I.N.P.G. , 46 AVENUE FELIX VIALLET , GRENOBLE.

ABSTRACT.

In this paper, our interest concerns the robustness of seven pitch detector algorithms (PDA's) that we have implemented. The PDA's are constituted of four temporal methods (*Amdf, Dubnowsky, Sondhi, modified Rabiner*), two spectral methods (*Sift, Cepstral*) and a mixed method (*Modified Ambiguity*). The experimental study has been carried out on a large Arabic data basis of microphonic and electroglottographic signals. The former signals have been corrupted with seven pseudo-gaussian noise levels (7 different RSB) before being analysed while the later signals have been used as a reference. New improvements to the standard algorithms and new techniques of smoothing have been introduced to perform well with these algorithms in noisy speech. For every PDA and every RSB, eight kind of errors have been measured and compared (No errors, fine errors, Gross errors, V/UV errors, mean of errors, standard deviation of errors, error on the mean of errors, absolute errors).

1. INTRODUCTION.

Une étude comparative des performances de sept détecteurs de pitch est effectuée ici dans de sévères conditions de bruitage.

Les méthodes étudiées sont: l'***Amdf*** (Rabiner, 1976), L'autocorrelation à codage *sign* dite méthode de ***Dubnowsky*** (Rabiner, 1976), l'autocorrelation à codage clc dite méthode de **Sondhi** (Rabiner, 1976) (Sondhi, 1968), la méthode ***Cepstrale***(Rabiner, 1976), la méthode ***Sift*** (Rabiner, 1976) (Markel, 1972), la méthode d'***Ambiguité modifiée*** (Selouani, 1992) (Boudraa, 1994) et enfin la méthode parallèle dite de ***Rabiner*** (Rabiner, 1976) (Boudraa, 1993).

Toutes ces méthodes sont testées dans les mêmes conditions de bruitage et admettent, pratiquement, les mêmes pré et post-traitements.

La base de données traitée est constituée de 10 phrases arabes phonétiquement équilibrées, d'environ trois secondes de parole chacune prononcées par six locuteurs dont trois masculins et trois féminins. Ces phrases ont été enregistrées simultanément avec le signal laryngographique pris comme référence mélodique. l'ensemble de soixante fichiers de parole est ensuite corrompu par un bruit blanc gaussien additif, pondéré à sept niveaux et dont les RSB respectifs sont ***96***, ***+12***, ***+6***, ***0***, ***-6***, ***-12***, et ***-18 dB***. Ce bruitage est appliqué à chaque fenêtre de 20ms.

Avant le début de la détection, le même prétraitement est effectué pour chaque méthode. Il est constitué d'un filtrage passe-bas à fréquence de coupure de 600 Hz, suivi du calcul de l'énergie moyenne de la fenêtre et du taux de passage par zéro(TPZ) correspondant, dans le but de localiser respectivement les zones de silence et de non voisement.

Notons que la fenêtre d'analyse a été choisie de 20ms avec un recouvrement de 50%, pour une fréquence d'échantillonnage de 10 Khz.

2. METHODES DE DETECTION.

Les sept PDA's implantés peuvent être classés suivant trois catégories: temporelles, spectrales et mixtes, et ceci, selon leur nature algorithmique et leur principe théorique.

2.1.Méthodes temporelles.

- La méthode ***Amdf***
- La méthode de ***Dubnowsky***
- La méthode de ***Sondhi***
- La méthode de ***Rabiner modifiée.*** Il s'agit de notre variante de la méthode de ***Gold & Rabiner*** (reposant sur la position des pics et des vallées). L'amélioration que nous avons apportée à cette méthode a été détaillée en (Boudraa, 1993).

2.2. Méthodes spectrales.

- Méthode ***Sift***: c'est la méthode introduite par MARKEL dans (Markel, 1972) et dont la détection est appliquée au résiduel de prédiction linéaire, après filtrage inverse d'ordre 4. La détection se fait sur l'autocorrélation de ce résiduel.
- Méthode ***Cepstrale*** (Rabiner, 1976).

2.3. Méthode mixte.

Nous avons considéré ici une méthode dénommée ***Ambiguité modifiée*** et qui englobe une

analyse temporelle et une analyse fréquentielle. Elle est basée sur les 3 critères de l'ambiguité et sur un codage spécifique simplifiant le calcul de l'ambiguité que nous avons défini dans (Selouani, 1992) (Boudraa, 1994).

3. POST TRAITEMENTS.

Les résultats obtenus à la fin de la détection sont raffinés grâce à l'application de trois post-traitements: un suivi dynamique, des filtrages logiques (continuité, anti-double-période, anti-pic), et des lissages divers (médian fin, médian gros, d'ilôt et de bordure)

Le suivi dynamique consiste en un choix d'un chemin optimal se faisant sur au plus un treillis de trois candidats variant au cours du temps. Le chemin optimal sera, à notre sens, celui qui recevra le plus faible coût (distance minimale).

Le test de continuité découle de l'inertie physico-acoustique de la parole, interdissant toute variation brutale de F_0 dépassant 20% durant 20ms.

$$\frac{\Delta F_0}{F_0} < 20\ \%$$

Le lissage médian fin est un lissage noté médian5. C'est un moyennage à pondération centrale et sa formule est donnée par:

$$F_{0lissée}[i] = \sum_{j=-2}^{+2} \alpha_j F_0[i+j]$$

αj suit une pondération gausienne, i est le rang de la fenêtre. Ce lissage n'est appliqué que si l'écart entre la valeur originale et la valeur lissée est inférieur à 5%. Ce lissage a été mis au point pour améliorer le taux des plages dites sans erreurs:

$$\text{Si} \quad \frac{|F_{0lissée} - F_0|}{F_0} < 5\% \quad \text{alors} \quad F_0 = F_{0lissée}$$

Le lissage médian gros est un lissage médian5, qui n'est appliqué que si l'écart entre la valeur orginale et la valeur lissée est supérieur à 20%. Ce dernier lissage a réduit considérablement le taux des grosses erreurs dans notre cas.

Le lissage de bordure est utilisé là où le lissage médian est inapplicable à savoir dans les zones de transitions.

Le filtrage anti-pics est utilisé dans les cas où le premier harmonique prédomine sur le pitch. Pour y remédier, on procède à une correction appropriée.

Le filtrage anti-double période est utilisé là où les pics du signal original sont irréguliers dans de larges proportions.

4. RESULTATS.

Les pré-traitements et les traitements (détection) ont été appliqués à tous les enregistrements pondérés par les 7 niveaux de bruits précédents. Les résultats obtenus par chaque méthode et pour chaque RSB, sont comparés par rapport aux références électroglottographiques, pour chiffrer les huit types d'erreurs. Nous donnerons quelques unes:

E_r : l'erreur relative sur F_0 est calculée par:

$$\frac{\Delta F_0}{F_0} = \frac{(F_0 - F_{0\text{REF.}})}{F_{0\text{REF.}}}$$

μ : L'espérance statistique des erreurs est calculée par la formule suivante:

$$\mu = \frac{1}{N} \sum_i \left(\frac{\Delta F_{0i}}{F_{0i}}\right) P\left(\frac{\Delta F_{0i}}{F_{0i}}\right)$$

où $P\left(\frac{\Delta F_{0i}}{F_{0i}}\right)$ représente la loi statistique de $\frac{\Delta F_{0i}}{F_{0i}}$

S.E: taux des sans erreurs (E_r<5%). Il donne une indication sur la finesse du détecteur et il est d'autant plus grand que le détecteur est fin et selectif. Un détecteur est de bonne qualité si ce taux est supérieur à 95%.

E.F: taux des erreurs fines ($5\% \le E_r \le 10\%$)

G.E: taux des grosses erreurs ($E_r > 10\%$). Un détecteur sera d'autant plus précis si ce taux est inférieur à 5%.

EMP : Erreur sur la moyenne du pitch,

ROB: coefficient de robustesse. Nous l'avons élaboré dans notre laboratoire, c'est un paramètre qui symbolise la résistance du détecteur vis-à-vis du bruit. On le définit comme étant le rapport du taux des trames sans erreurs à -6 dB , sur le taux des trames sans erreur à un RSB infini (sans bruit):

ROB= (SE à -6 dB) / (SE à ∞ dB)

Un détecteur est insensible au bruit si Rob est très proche de 1. Ainsi, on voit, sur la figure 1, que le taux des sans erreurs (E_r<5%) est sensiblement le même pour toutes les méthodes, lorsque le RSB est élevé. Mais, il commence à diverger dès que l'on passe en dessous de 0 db. L'***Ambiguité*** présente alors le meilleur taux suivie du PDA de ***Dubnowsky***.

Sur la figure 2, on remarque un maximum d'erreurs fines autour de -12 dB pour toutes les méthodes. Cette courbe représente le flux des erreurs de la zone des sans-erreurs vers la zone des grosses-erreurs. Le début de décroissance des erreurs fines ($5\% \le E_r \le 10\%$) indique alors la défaillance du détecteur. Toutes les méthodes ont une défaillance à environ -12 dB, sauf l'***Ambiguité*** dont la défaillance est retardée jusqu'à -18 dB.

La figure 3 représentant le taux des grosses erreurs (E_r>10%) confirme également que la courbe la plus basse est celle de l'***Ambiguité***, suivie de celle de ***Dubnowsky*** et ceci dès que le RSB descend en dessous à -6dB.

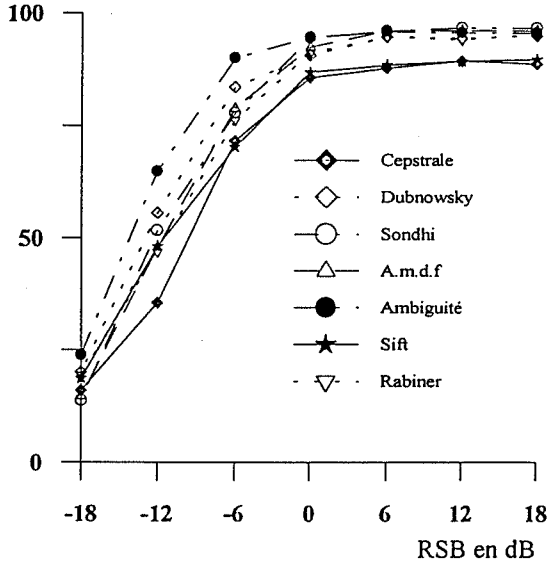


Figure 1: Taux (%) des Sans Erreurs des 7 Méthodes pour différents RSB.

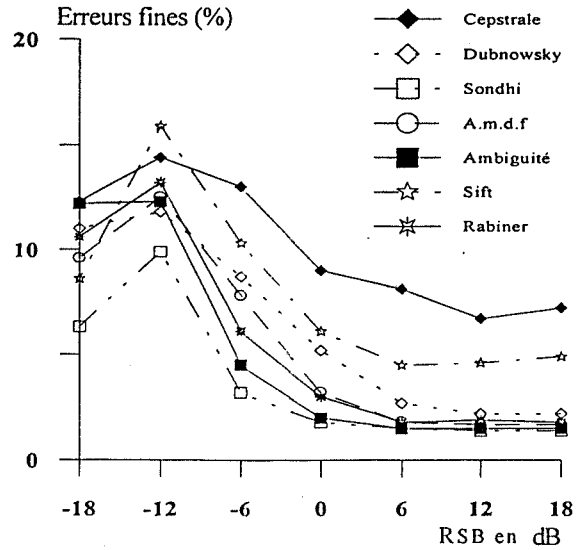


Figure 2: Erreurs Fines des 7 méthodes pour différents RSB.

Les méthodes spectrales (***Sift*** et ***Cepstrale***) possèdent les plus grands taux d'erreurs. Cependant, dans le cas non bruité, la méthode la plus précise est celle de ***Sondhi***.

Sur la figure 4 on distingue trois types d'espérances: positive, négative ou proche de zéro. L'***Ambiguité***, l'***Amdf*** et ***Sondhi*** possèdent des espérances proches de zéro. Ceci explique l'équitable répartition des erreurs par rapport à zéro. Pour le PDA ***Sift***, l'espérance s'écarte brusquement de zéro et demeure positive à cause de l'influence des harmoniques. Les trois autres méthodes: ***Rabiner***, ***Dubnowsky*** et ***Cepstrale*** présentent une espérance négative non négligeable. Ceci exprime la sensibilité de ces méthodes aux formants et aux doubles périodes.

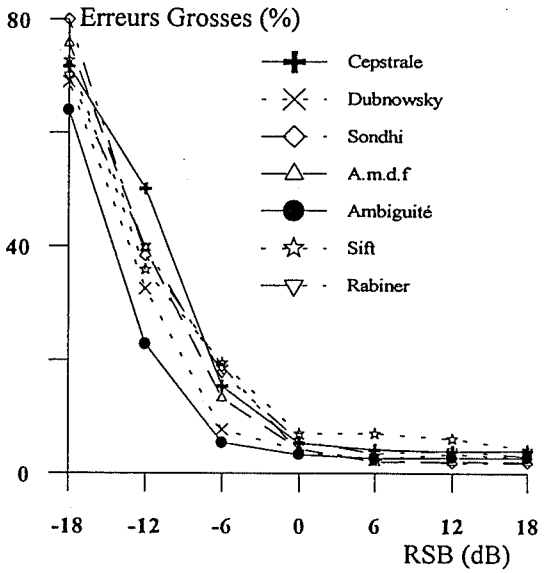


Figure 3: Erreurs Grosses des 7 Méthodes en fonction du RSB.

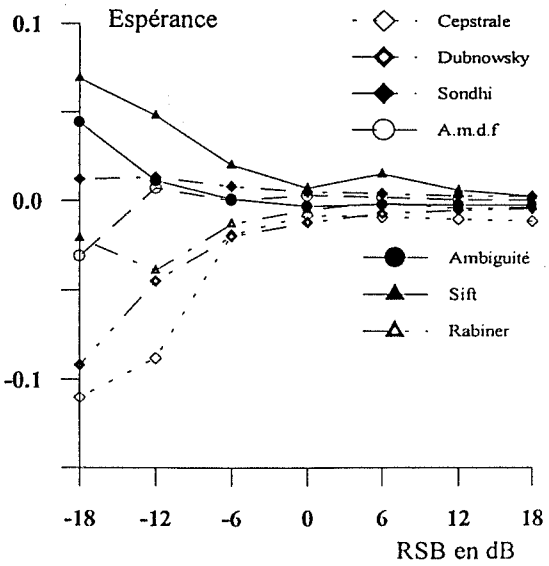


Figure 4: Espérance des erreurs des 7 méthodes pour différents RSB.

La figure 5 représente les erreurs sur la moyenne du pitch. L'ambiguité est la méthode la plus performante, suivie de celle de ***Dubnowsky*** pour les niveaux de RSB<0 dB. En dessous de , la méthode ***Sondhi*** se dégrade nettement, suivie de l'***Amdf***.

Nous remarquons, sur la figure 6, que l'***Ambiguïté*** présente un taux de robustesse très proche de l'unité (0.94) , ce qui confirme la très

bonne immunité de cette méthode contre le bruit blanc.

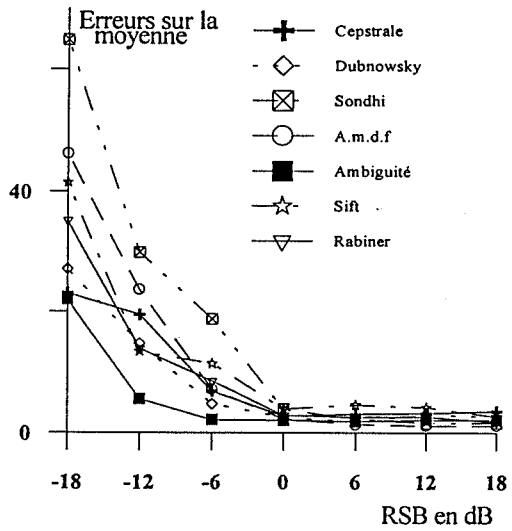


Figure 5: Erreurs sur la Moyenne du Pitch pour différents RSB.

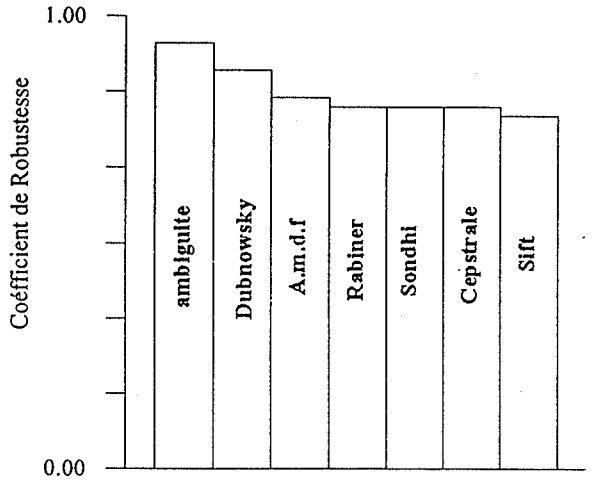


Figure 6: Coefficients de robustesse des 7 méthodes

La méthode de ***Dubnowsky*** présente une assez bonne robustesse avec un coefficient de 0.88. Les autres méthodes ont des coefficients de robustesse qui avoisinent 0.80 et se positionnent, alors, dans la classe des détecteurs peu robustes. Ainsi, à un niveau de bruit avoisinant les -6 dB, la méthode d'***Ambiguïté*** et ***Dubnowsky*** sont les seules méthodes pouvant être efficacement utilisée dans ces conditions de bruitage.

Donc le codage SIGN s'impose bien en milieu bruité.

5. CONCLUSION.

Les traitements faits sur la base de données sus-citée ont donné des résultals expérimentaux assez significatifs. Ainsi, du point de vue précision sur des signaux non bruités, le classement des différents PDA's, par ordre décroissant en précision, a donné le résultat suivant: ***Sondhi, Amdf, Ambiguité, Dubnowsky, Rabiner, Sift, Cepstrale.***

Par contre, en milieu bruité et pour des RSB inférieurs à +6dB, l'ordre en précision a complètement basculé en faveur de l'ambiguité exclusivement: ***Ambiguité, Dubnowsky, Rabiner, Sift, Cepstrale, Amdf*** et ***Sondhi.***

Par ailleurs, les deux méthodes utilisant le codage non linéaire ***Sign*** ont complétement émergé au dessus de toutes les autres méthodes dès que l'on descend en dessous de -6dB. Nous en tirons une conclusion concrète sur le bon filtrage du bruit blanc par le codage signe, en plus du gain en rapidité qu'il proccure.

6. BIBLIOGRAPHIE.

Boudraa M. (1993) "*Mélodimètre amélioré de Gold Rabiner*", actes du séminaire de l'Institut d'électronique, USTHB, Alger.

Boudraa M., Sayoud H., Boudraa B., Guérin B. (1994) "*PDA à ambiguité modifiée. Robustesse sur la parole corrompue*", Proc. of the International Conference on Signal and Systems, ICSS'94, Alger, n°2, IV-16 à IV- 20.

Markel J.D. (1972) "*The SIFT algorithm for fundamental frequency estimation*", IEEE, Trans. on ASSP, n°5, 367-377.

Rabiner L., Cheng M.J., Rosenberg A.E., McConegal C.A. (1976) "*A comparative performance study of several pitch detection algorithms*", IEEE trans. on ASSP, n° 24, 399-417.

Selouani S.A., Boudraa M., Boudraa B., Guerin B. (1992), "*Extraction de la fréquence fondamentale du signal de parole par la fonction d'ambiguité. Nouvelle approche*". Séminaire SFA-GCP, "*Traitements et représentation du signal de parole*", Le Mans.

.Sondhi M.M. (1968) "*New méthods of pitch extraction*". IEEE trans. on audio electroacoustic, n°. 16, 262-266.

UNE NOUVELLE ESTIMATION DE COEFFICIENTS CEPSTRAUX POUR LA RECONNAISSANCE AUTOMATIQUE DE LA PAROLE

Hubert Wassner *, Gérard Chollet * **

*IDIAP, CP 592, 1920 Martigny, Suisse, ** CNRS URA-820, ENST, PARIS

wassner@idiap.ch, chollet@idiap.ch

ABSTRACT

This work is about improving recognition rate by calculating noise resistant cepstral coefficients. It varies in the estimation of th short term power spectrumand the weighting of this spectrum. There are many ways to obtain a spectrum out of a signal which differs in the method itself (Fourier, Wavelets,...), and in the normalisation. We show here that we can obtain noise resistant cepstral coefficients, for speaker independant speech recognition.

The continuous speech recognition system is based on a continous whole word hidden Markov model.

An error reduction rate of approximately 50% is achieved. Moreover tests have been made that show that these new parametrisations allows smaller databases to obtain similar results, halving the training database have small effects on recoginition rates (which is not the case with classical 'mfcc').

1. INTRODUCTION

Il s'agit d'optimiser l'estimation des coefficients cepstraux dans le cadre d'un système de reconnaissance de la parole continue indépendante du locuteur (via HMM). Ces adaptations interviennent dans la première partie du calcul des cepstres: l'obtention d'une image temps-fréquence à partir du signal de parole.

On montre qu'il y a des gains significatifs à obtenir , par rapport à la méthode la plus couramment utilisée, en portant un soin particulier à la méthode temps-fréquence et sa "normalisation". Les améliorations sont de deux ordres:

- Une réduction de 50 % du taux d'erreur.
- Une réduction de 50 % de la taille des données d'apprentissage nécéssaire.

Les coefficients les plus utilisés en traitement automatique de la parole, sont les MFCC (Mel Frequency Cepstral Coefficients). La première partie est un rappel de la méthode classique de calcul de ces coefficients. La deuxième partie est consacrée au différentes améliorations proposées et réalisées dans cette étude. Enfin une troisième partie détaille les bases de données utilisées, les expériences les plus intéressantes, les résultats obtenus et leurs interprétations.

2. Rappel sur l'estimation des MFCC

Les MFCC servent à réduire la quantité de données pour représenter un signal de parole. Ils décrivent l'enveloppe spectrale de ce signal et forment ainsi un pré-traitement intéressant pour la reconnaissance de la parole indépendante du locuteur. De nombreuses études ont montré l'intérêt de calculer les coefficients cepstraux suivant une échelle Mel en fréquence. Voici en 2 étapes une description du calcul classique des MFCC (Rabiner, 1992):

1- Log de l'énergie en sortie d'un banc de filtres.

2- Transformé en cosinus de la log-énergie en sortie du banc de filtres.

L'étape 1 peut être simulée par :

1a- Le calcul du spectre de puissance.

1b- Intégration dans un banc de filtres.

Le schéma suivant (fig. 1) résume les étapes précédemment citées pour l'obtention de cepstres à partir du signal de parole.

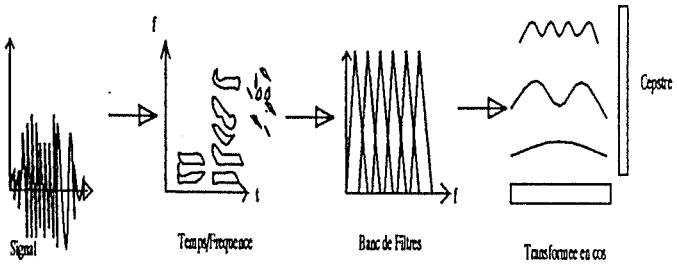


fig. 1: Du signal aux cepstres

2 Coefficients Cepstraux "Ameliorés"

Le spectre de puissance est le plus souvent estimé par FFT mais d'autres transformations temps/fréquence sont utilisables. Par exemple une transformation en ondelettes, permettant

un meilleur compromis temps/fréquence, peut tout à fait permettre d'obtenir le spectre desiré. Nous avons utilisé une transformée en ondelettes définie par (Unser, 1994):

$$Wx(t,a) = 1/\sqrt{a} \sum_{T=-\infty}^{T=+\infty} x(t+T)\, g(a,T)\, e^{-i\,\Omega\, T/a}$$

a est le facteur d'échelle, relié à la fréquence par: $f = \Omega/(2\,\pi\,a)$.

$g(a,t) = 1/\sqrt{(2\,\pi)}\, e^{-t^2/(2\,a^2)}$ est une fenêtre dont la taille dépend de a.

Un choix judicieux des facteurs d'échelle a permis de simuler un banc de filtres dont les fréquences sont réparties sur une échelle de fréquence quelconque (Mel en particulier).

La transformée de Fourier à court terme ((Cohen, 1989), (Riley, 1989)):

$$Sx(t,f) = \sum_{s=-\infty}^{s=+\infty} h(s)\; x(t-s)\, e^{-2\,i\,\pi\,f\,s}$$

(ou h(s) est une fenêtre, par exemple une gaussienne) peut être vue comme une transformée en Ondelettes, avec $a = \Omega/(2\,\pi\,f)$ et $g(a,t) = h(t)$.

La seule différence réside dans le fait que la taille de la fenêtre est constante pour Fourier et variable pour les Ondelettes. La notion d'échelle dans les Ondelettes est simplement un changement de variable dans la transformée de Fourier.

Dans le calcul traditionel des cepstres on utilise une transformation log pour modifier les énergies, que ce soit avant ou après l'intégration par banc de filtres. On verra plus tard qu'il est plus intéressant, dans notrer cas, de placer cette "normalisation" (qui est plutôt une transformation) avant le banc de filtres (cf. fig. 3 et 4 pour voir la sortie du banc de filtres). De plus on constate que dans certaines situations, enregistrements bruités par exemple, cette fonction de normalisation fait trop ressortir le bruit de fond, en laissant "trop" de dynamique dans les basses énergies que l'on peut attribuer au bruit de fond. On peut donc penser à utiliser d'autres fonctions pour normaliser le spectre (fig. 2).

3 Expérience et Résultats

Le protocole des tests a été le même pour toutes les expériences, seule la paramétrisation du signal a été différente. On effectue une initialisation des modèles puis une ré-estimation (programmes HInit et HRest d'HTK (HTK, V1.5)). Deux de ces expériences (pour avoir des résultats de référence) ont été effectuées en paramétrisant le signal avec le programme HCode d'HTK, toutes les autres on été faites en utilisant un programme développé pour l'occasion (HCepstre). Les modèles de Markov utilisés sont des modèles gauche-droite représentant des mots.

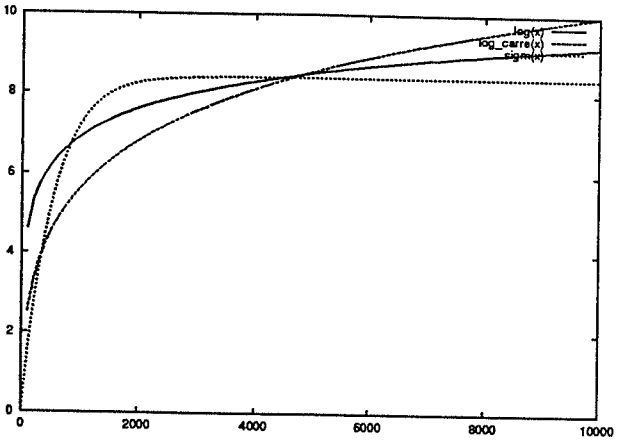


fig.2: Fonctions de normalisation de l'énergie.

Les options qui ont été choisies par défaut pour toutes les expériences sont les suivantes:

- Echelle Mel.
- Nombre de cepstres : 12
- Nombre de filtres : 24
- Cepstral liftering: 22
- Pré-accentuation: 0.98
- Fréquence d'analyse 10 ms
- Taille de la fenêtre 25 ms

(Fourier) variable pour ondelettes.

Chaque vecteur MFCC est composé de 12 coefficents cesptraux, on y ajoute l'énergie, la dérivée et l'accélération du vecteur de 13 composantes ainsi créé. On a donc au total un vecteur de 39 composantes pour chaque trame.

Ces paramètres sont ceux qui sont proposés dans un document d'aide du logiciel HTK (HTK, 1993). Rien ne prouve vraiment que ce sont les meilleurs dans toutes les situations mais c'est un ensemble de paramètres "qui à fait ses preuves". C'est pourquoi nous les avons pris comme référence dans nos expériences.

Nous avons utilisé trois bases de données, deux extraites de la base "Polyphone" et une extraite de la base "Computer95". Ces deux bases ont été crées par l'IDIAP et les telecom PTT Suisses, il s'agit de parole passée à travers le réseau téléphonique Suisse. Dans les trois cas, les extraits sont les chiffres de zéro à neuf prononcés en français. Dans les extraits de PolyPhon, il y a en plus les mots "étoile" et "dièse". La base PolyPhon est faite par des gens qui appellent de chez eux, donc avec un bruit de fond assez faible. Par contre la base Computer95 a été enregistrée dans le hall d'exposition de "Computer95", forum d'informatique annuel à Lausanne. Les trois extraits comportent chacun des locuteurs différents, il y a au total 1303 locuteurs qui ont prononcé 9296 mots. La répartition se faisant selon le tableau No 1.

Tableau 1: Taille des bases de données.

	Nbr. loc.	Nbr. mots
Apprentissage (Polyphone)	498	2962
Evaluation (Polyphone)	429	2574
Test (Computer95)	376	3760

Le tableau No 2 résume les résultats (en %) des expériences les plus intéressantes.

On constate tout de suite qu'il y a des **gains non négligeables** par rapport à l'expérience de référence (1). On se rend compte que les gains semblent être plus importants sur la base "Computer95". En fait si on y regarde de plus près, on se rend compte que dans les meilleurs cas et quelque soit la base **on divise le pourcentage d'erreurs par deux.**

On note que le simple déplacement de la fonction Log dans la chaîne de traitement augmente sensiblement les performances. On a donc choisi, pour toutes les autres expériences, de placer la transformation de l'échelle d'énergie du spectre avant le banc de filtres. Plus d'information au sujet de ce choix peuvent être trouvées dans (Wassner, 1996).

Dans la méthode mettant en oeuvre une transformée de Fourier on a remplacé le log par une sigmoïde, on constate que les meilleurs résultats concordent sur la base PolyPhone et la base Computer95, selon le coefficient de la sigmoïde, ce qui est appréciable d'un point de vue pratique pour de futures applications, et qui semble indiquer qu'il existerait une fonction optimale indépendante de la base utilisée.

Les expériences 12 et 13 montrent la possibilité offerte par la nouvelle paramétrisation de faire **un apprentissage 2 fois plus rapide,** tout en perdant peu de pourcentage de reconnaissance. Même avec une base divisée par deux on arrive avec la nouvelle paramétrisation à **des résultats bien supérieurs aux MFCC classiques.**

On constate qu'une courbe de modification d'énergie est intimement liée à la transformation TF qui la précède. Par exemple, on peut noter une différence de 3% sur PolyPhone et de 10 % sur Computer95 entre les couples (MWCC,Log^2) et (MFCC,log^2).

Si on effectue une re-estimation en contexte (programme HERest d'HTK) avec 390 locuteurs supplémentaires, extraits de la base Polyphon, soit 2340 mots, on obtient les résultats (en %) suivants (tableau No 3).

la nouvelle paramétrisation reste meilleure que la paramétrisation classique, mais la **différence s'amenuise,** en tout cas sur la base Polyphone. **La différence reste toujours largement significative sur la base Computer95.** Mais surtout on constate que la re-estimation en contexte n'apporte presque plus rien en pourcentage de reconnaissance avec la nouvelle paramétrisation, **cette phase semble donc devenir ici presque inutile.** Cela permettrait de faire des économies en bases de données.

4. Conclusion

Il semble maintenant clair que le calcul de Cepstre, tel qu'il est le plus répandu n'est de loin pas une solution optimale. Tout en gardant la même structure, on peut avoir un pré-traitement bien plus efficace dans n'importe quelle condition (bruité ou non bruité). Surtout **il est possible d'effectuer des apprentissages avec des bases deux fois plus petites.** Cette caractéristique est importante puisqu'une des difficultées en TAP est la collection coûteuse de bases de

données. Nous avons déjà trouvé plusieurs modifications permettant de **diviser par deux le nombre d'erreurs**. Ces modifications seraient peut-etre encore plus efficaces en affinant leurs paramètres. **Une recherche automatique** de ces paramètres, quoique non triviale, **semble être possible**. Une borne inférieure des gains possibles en terme de diminution de l'erreur est une diminution de moitié, ce qui est déjà intéressant. Maintenant seule l'expérience d'une telle recherche nous donnera la diminution exacte d'erreur que l'on peut obtenir.

5. Bibliographie

Lawrence Rabiner, Biing-Hwang Juang, (1993) "Fundamentals of Speech Recognition", Prentice Hall signal processing series, Alan V. Oppenheim, series editor.

Michael Unser, (1994) "Fast Gabor-Like Windowed Fourier and Continuous Wavelets Transforms", IEEE Signal Processing Letters, vol 1,No 5, may.

Michael Unser, Akram Aldroubi, Steven J. Schiff, (1994) "Fast Implementation of the Continuous Wavelet Transfom with Interger Scales", IEEE transactions on acoutics, speech, and signal processing, vol 42, No 12 , december.

Léon Cohen, (1989) "Time-Frequency Distribution -A Review" Proc. IEEE, vol 77, No 7, july.

Michael D. Riley, (1989) "Speech Time-Frequency Representations", Kluwer Academic Publishers.

HTK (Hidden Markov Model Toolkit V1.5)

"Using HTK to Design A Speaker-Independant Connected-Digit Recognition System." (1994) Entropic Research Laboratory, Washington DC, August.

Hubert Wassner, rapport IDIAP, (1996) "Etude sur la Paramétrisation du Signal en traitement automatique de la parole", IDIAP Janvier.

Charles K. Chui, (1992) "Wavelet Analysis and its Applications, vol 1: introdution to Wavelets" Academic Press.

Lawrence R. Rabiner, Bernard Gold, (1975) "Theory and Application of Digital Signal Processing", Prentice Hall.

"Fundamental of Speech synthesis and Speech Recognition" (1994) Edited by Eric Keller.

Thèse: Chafic Mokbel, "Reconnaissance de la Parole dans le Bruit': Bruitage/Debruitage", Telecom Paris 92 E 008.

Journées thématiques du GDR TdSI et colloque, (1994) "Temps-Fréquence et Multirésolution : théories modèles et applications" INSA Lyon 9-11 Mars.

Tableau 2: Résultats des différentes expériences.

	Paramétrisation	Polyphone	Computer95
1	MFCC de référence (HTK)	92.70	65.48
2	MFCC log avant banc de filtres	97.40	79.36
3	MFCC log après banc de filtres	91.18	62.13
4	MFCC Sigmoïde (60)	93.28	67.31
5	MFCC Sigmoide (30)	96.70	75.11
6	MFCC Sigmoïde (15)	96.27	80.88
7	MFCC Sigmoïde (10)	95.18	81.33
8	MWCC, $\Omega = 9$, log	96.89	80.40
9	MWCC, $\Omega = 11$, log	97.47	82.61
10	MWCC, $\Omega = 9$, $\log^2$	96.70	81.62
11	MFCC $\log^2$	93.59	71.86
12	MFCC (HTK) sur 45 % de la base	87.61	59.81
13	MWCC $\Omega = 11$, log sur 45 % de la base	96.15	80.16

Notes:
-MWCC: ce sont des MFCC calculés sur une image temps fréquence calculée par Ondelettes.
-log, sigmoïde, $\log^2$: sont les différentes fonctions de "normalisation" de l'énergie. En dehors des expériences 1, 3, 11, cette "normalisation" se situe avant le banc de filtres. Dans le cas de la sigmoïde, un coefficient est spécifie, il influe sur la pente de la sigmoïde.
-Les expériences 12 et 13 ont été effectuées avec un apprentissage sur environs 45 % de la base, ceci afin d'étudier la vitesse d'apprentissage.

Tableau 3: Résultats après re-estimation en contexte.

	Paramétrisation	Polyphone	Computer95
1	MFCC de réference (HTK)	96.81	76.54
9	MWCC, $\Omega = 11$, log	97.98	83.48

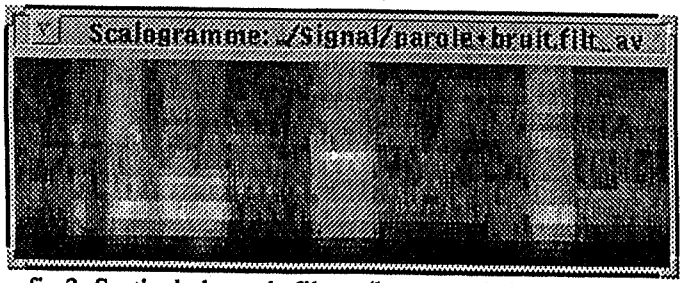


fig.3: Sortie du banc de filtres (log avant le banc de filtres)

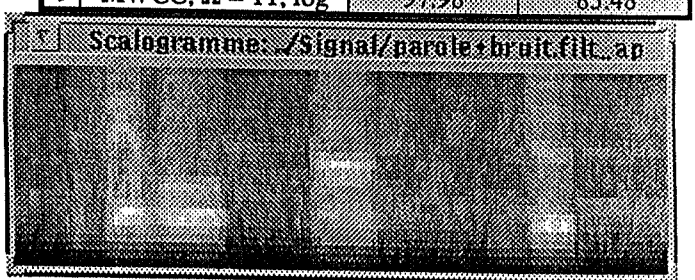


fig. 4: Sortie du banc de filtres (log après le banc de filtres)

JEP 96

RECONNAISSANCE

AVIGNON 10-14 JUIN 1996

RECONNAISSANCE AUTOMATIQUE DE LA PAROLE: MODÉLISATION OU DESCRIPTION?

Hervé Bourlard

Faculté Polytechnique de Mons, Mons, Belgium
International Computer Science Institute, Berkeley, CA 94704, USA
Email: bourlard@tcts.fpms.ac.be

ABSTRACT

In the field of automatic speech recognition research, it has become conventional to choose research directions in order to reduce the word error rate in standard evaluations with predefined reference databases. This strategy led to major improvements in speech recognition technology. However, this was achieved primarily by small improvements of the leading approach over a number of years. To some extent, this successful approach may have effectively optimized the systems for a seriously suboptimal model, in effect approaching a local minimum in the space of available techniques. Initially, this was exciting for all of us since this research led to clear improvements on challenging tasks. However, we also have the feeling that this research strategy led scientists to put too much emphasis on short term goals, limiting the risks and pushing further the limits of the description capabilities of the existing models, instead of extending their actual speech modelling capabilities. To illustrate our point, we present some alternative approaches deviating from the current dominant paradigm and that, we believe, could exhibit better modelling properties.

1. INTRODUCTION

Thus I saw that most men only care for science so far as they get a living by it, and they worship error when it affords them a subsistence.
— Johann Wolfgang von Goethe (1749-1832) —

Depuis ces quelques dernières années, il semble que, mis à part quelques modifications et améliorations marginales, la seule manière d'"améliorer" nos systèmes actuels de reconnaissance automatique de la parole basés sur la technologie des *modèles de Markov cachés* (HMM, pour "Hidden Markov Models") soit d'accroître la taille de nos bases de données d'entraînement, le nombre d'unités de parole élémentaires et le nombre de paramètres décrivant ces modèles. Il est vrai que ce type d'approche a permis aux scientifiques de développer des systèmes capables de reconnaître des vocabulaires de plus en plus étendus avec des performances de laboratoire assez impressionnantes. Malheureusement, cela c'est fait sans pour autant en améliorer la robustesse aux variations inhérentes aux conditions réelles d'utilisation et, souvent, sans vérifier que les conclusions obtenues sur de nouvelles bases de données restaient valables pour les bases de données plus anciennes et moins "compétitives".

Nous croyons également que cette méthodologie de recherche a conduit à des systèmes de reconnaissance de la parole se focalisant surtout sur la description du signal de parole plutôt que sur sa véritable modélisation, résultant ainsi en de mauvaises propriétés de généralisation. Dans cet ordre d'idées, nous discuterons ici de quelques techniques qui ont démontré expérimentalement (sur certaines bases de données de référence) qu'elles pouvaient conduire aux mêmes taux de reconnaissance (parfois meilleurs) que les meilleurs des systèmes HMM "classiques" existants, tout en utilisant un nombre nettement plus faible de paramètres et nécessitant beaucoup moins de temps de calcul. Ce papier n'a certainement pas la prétention de présenter la solution au problème de reconnaissance de la parole, mais simplement de discuter de certaines alternatives (basées sur des considérations théoriques et expérimentales solides) à l'approche dominante utilisée actuellement. Parmi ces alternatives, nous discuterons plus précisement des récents développements et résultats obtenus dans différents domaines tels que: (1) systèmes hybrides HMM et *réseaux de neurones artificiels* (ANN, pour "Artificial Neural Networks"), (2) nouveaux modèles plus discriminants et basés sur un nouveau type d'automate stochastique et (3) une meilleure utilisation de l'information temps-fréquence, en vue notamment d'une meilleure robustesse aux perturbations non observées sur l'ensemble d'entraînement.

2. DISCUSSION GENERALE

Lorsque Copernic (1473-1543) introduisit son modèle héliocentrique du système solaire, en faisant l'hypothèse de trajectoires circulaires des planètes autour du soleil, ce modèle conduisit initialement à des erreurs largement supérieures à celles obtenues en utilisant le modèle bien connu et relativement bon

(bien que faux) de Ptolémée (2ème s. apr. J.-C.) pour lequel le soleil et les planètes étaient supposés tourner autour de la terre selon des orbites très compliquées. Heureusement pour l'Humanité, ces erreurs de modélisation ne découragèrent pas Copernic. Cependant, pour convaincre ses collègues de la validité de son modèle, Copernic dut appliquer une "astuce d'ingénieur" empruntée d'ailleurs au système pragmatique de Ptolémée: il postula l'existence d'épicycloïdes, c'est-à-dire de petites sous-révolutions des planètes lors de leur mouvement le long de leur orbite de révolution principale. Cette modification du modèle de base permit de rendre le modèle de Copernic compétitif par rapport au modèle de Ptolémée. Evidemment, afin d'éviter tout problème avec ses confrères scientifiques ainsi que l'Eglise (on comprends facilement pourquoi en pensant à Galilée), Copernicus n'hésita pas à reconnaître que son modèle, contrairement à celui de Ptolémée, ne reflétait pas la vérité mais avait simplement l'avantage d'être plus élégant et plus simple à traiter. Comme nous le savons, Copernicus n'était pas très loin du "vrai" modèle; il fallut cependant attendre Kepler (1571-1630) pour l'introduction des orbites elliptiques et l'abandon des épicycloïdes.

De façon un peu similaire, les pionniers des méthodes "modernes" de reconnaissance automatique de la parole (Jelinek, 1975; Baker, 1975) ont développé des modèles (HMMs) dont la mise en oeuvre théorique et pratique est relativement simple. Ces nouveaux modèles, que nous pourrions comparer au modèle de Copernic, ont permis d'améliorer les systèmes précédemment disponibles et basés sur une simple comparaison (DTW) de formes (une ou plusieurs prononciations de chaque mot étant mémoriées comme des formes de référence). Cette dernière technique pourrait être comparée au modèle de Ptolémée. Comme pour le modèle de Copernic, il fallut cependant attendre un certain temps avant que la technique HMM ne soit acceptée et démontrée meilleure que l'approche DTW[1].

Durant ces 10-20 dernières années, la plupart des recherches en reconnaissance de parole se sont focalisées sur l'amélioration de ces modèles HMMs et ont conduit à des systèmes relativement sophistiqués dont les performances en laboratoire peuvent être impressionnantes. Il faut cependant admettre que, en grande partie, ceci ne résulte pas d'une amélioration du modèle de base, mais simplement d'une meilleure maîtrise de celui-ci et de son utilisation optimale d'ailleurs poussée à l'extrême. Evidemment, on peut toujours dire qu'aussi longtemps que l'on peut augmenter les performances d'une technologie existante on n'a pas atteint les limites du modèle. De nombreuses personnes (dont l'auteur de ce papier) ne croient cependant pas que tel est le cas, et cela pour plusieurs raisons:

1. Ce n'est pas parce que l'on est capable de traiter des problèmes de plus en plus compliqués (par exemple des vocabulaires de plus en plus larges) que l'on améliore le modèle. Il faut éviter de confondre le fait de rester actif avec le fait d'être productif et d'améliorer sa compréhension du problème. Il serait en effet bon de vérifier que les conclusions et améliorations obtenues sur de nouvelles tâches "plus difficiles" restent valables pour les bases de données plus anciennes et moins "compétitives".

2. Il n'est pas clair que l'amélioration des systèmes obtenue sur les bases de données de laboratoire traitent des véritables problèmes rencontrés dans les conditions réelles de fonctionnement. Il est en effet surprenant que malgré toutes les "améliorations significatives" de la technologie HMM telles que rapportées sur ces bases de données, il soit toujours impossible de reconnaître de façon robuste deux mots de vocabulaire sur ligne téléphonique, ou tout simplement de faire la différence entre la parole et le silence (ou, plus récemment, entre parole et musique).

3. Pour revenir brièvement à notre méthaphore astronomique, nous pourrions dire que les "améliorations" actuelles sont de l'ordre des épicycloïdes, réduisant le taux d'erreurs sur quelques observations, en compliquant le modèle initial et en réduisant ses propriétés de généralisation. Malgré les grands volumes des données d'entraînement actuellement traités, il est en effet bon de se rappeler que cela ne représentera jamais qu'un très faible échantillon de la variabilité réelle du signal de parole et des perturbations possibles.

Finalement, beaucoup s'accordent à dire que les modèles HMM de base actuellement utilisés pour la reconnaissance automatique de la parole n'ont plus été améliorés (du moins de façon significative) depuis ces 10 dernières années (les plus critiques diraient ces 20 dernières années). Dans un certain sens, ce qui a cependant été amélioré sont les capacités des modèles HMM à *décrire* les données plutôt qu'à les *modéliser*. Le terme "description" est défini ici comme le fait de simplement mémoriser les données d'entraînement, en augmentant le nombre de paramètres des modèles et en divisant les classes de base en sous-classes (triphones) sur lesquelles des statistiques simples peuvent être évaluées. Par contre, le terme "modélisation" fait référence à un processus d'unification qui pourrait représenter les données sous forme d'un structure de relativement faible dimension et pour laquelle les différentes sous-classes ne seraient que

[1] On note que jusque dans les années 1985 certains papiers de scientifiques de renom montraient que la technique DTW était supérieure à l'approche HMM.

des réalisations particulières de la même classe, conditionnée sur certaines variables pouvant affecter ces réalisations.

Nous croyons donc qu'il sera nécessaire de trouver un meilleur formalisme de reconnaissance de la parole. Idéalement, un Kepler (ou un groupe de Keplers) devrait développer l'équivalent des orbites ellyptiques pour la reconnaissance de la parole, c'est-à-dire un modèle (peut-être pas très loin des HMMs) qui reprendrait toutes les propriétés fondamentales de la reconnaissance de la parole avec de meilleures propriétés de généralisation. Jusqu'à présent, nous avons seulement des systèmes de type Copernic, c'est-à-dire basés sur des modèles reprenant beaucoup des caractéristiques nécessaires mais qui requièrent toujours trop d'épicycloïdes pour conduire à des résultats vraiment satisfaisants.

3. APPROCHE STANDARD...

... ou "Tout ce qu'il faut savoir pour construire son propre système de reconnaissance de la parole."

Errors using inadequate data are much less than those using no data at all".
— Charles Babbage (1792-1871) —

Dans cette section, nous résumons brièvement les caractéristiques principales des meilleurs systèmes "standards" de reconnaissance de la parole basés sur la technologie HMM.

Bien que les systèmes à petits vocabulaires utilisent encore souvent des modèles de mots, la plupart des systèmes de reconnaissance actuels utilisent des modèles de phonèmes. Dans ce cas, les meilleurs modèles HMM de phonèmes sont généralement basés sur une topologie à 3 états strictement gauche-droite. Il a effectivement été montré à plusieurs reprises que ce type de modèle conduit souvent aux meilleures performances. L'explication souvent donnée est que les deux états extrêmes capturent les effets de coarticulation alors que l'état central modélise la partie stationnaire du phonème. Bien que très élégante, il est cependant facile de vérifier que cette hypothèse n'est pas vérifiée en pratique. En effet: (1) la variance des observations associées aux états extrêmes n'est pas plus élevée que celle de l'état central, (2) la segmentation obtenue ne vérifie que très rarement cette hypothèse. De plus, si l'on permet le saut de l'état central, les performances du système diminuent, bien que le modèle soit plus général (sans accroissement significatif du nombre de paramètres).

Un autre résultat rarement publié est que les probabilités de transition ne jouent pratiquement aucun rôle sur les performances des systèmes de reconnaissance HMM[2]. Ces probabilités sont donc généralement abandonnées au profit d'une modélisation explicite de la durée de chaque phonème. Plusieurs méthodes ont été développées et comparées. Il semble cependant que la meilleure solution (comme souvent en reconnaissance de la parole) est également la plus simple et consiste à imposer un durée minimum fonction du phonème (simplement en répétant certains états qui ne peuvent être sautés... en quelque sorte une légère généralisation du modèle à 3 états strictement gauche-droite).

Ces modèles de phonèmes sont alors concaténés selon des règles phonologiques (voir prononciations multiples plus bas) de façon à obtenir des modèles de mots et de phrases utilisés pour la reconnaissance. Dans les meilleurs systèmes, les phonèmes sont complémentés par des modèles de diphones, triphones, et parfois quadriphones et quintaphones, qui sont combinés entre eux pour lisser les estimateurs des densités de probabilités. Il est bon de noter ici que ces modèles "dépendant du contexte" ne sont pas des modèles de séquences de phonèmes (contrairement aux diphones utilisés en TTS, alors définis comme des segments allant du milieu d'un phonème au milieu du phonème suivant). Dans le cas de la reconnaissance de la parole, ces unités de parole sont simplement introduites pour remédier artificiellement à notre manque de compréhension du problème et à l'incapacité de nos modèles d'extraire et de modéliser la variabilité du signal. Ce problème pourrait être du à différents facteurs tels que: (1) hypothèses trop restrictives sur la forme des densités de probabilités (généralement gaussienne), (2) mauvaise représentation du signal et (3) modèle non adéquat. Bien que beaucoup d'autres facteurs peuvent être en jeu, nous croyons (comme beaucoup d'autres) que la représentation du signal ainsi que la qualité du modèle (autre que simplement son nombre de paramètres) pourrait être les facteurs les plus importants.

Dans le cas de la reconnaissance de parole continue, plusieurs types de grammaires ont été utilisées pour contraindre le système à reconnaître des phrases syntaxiquement correctes. Ce papier ne discutera pas de cet aspect de la reconnaissance de la parole, mais il est clair que ce domaine, bien qu'important, soit encore largement sous-exploité. Dans la plupart des systèmes actuels, la façon d'utiliser ces grammaires est contrainte par la méthodologie de décodage basée sur les HMMs. De ce fait, la grammaire est souvent utilisée simplement dans le but de prédire les mots qui peuvent suivre un certain nombre de mots précédents, et cela sur base de la probabilité $p(W_k|W_{k-1}, W_{k-2}, \ldots, W_{k-N})$ d'un mot particulier W_k étant donné la séquence $\{W_{k-1}, \ldots, W_{k-N}\}$ des N mots précédents. Actuellement, les meilleurs systèmes utilisent conjointement des estimateurs d'unigrammes $p(W_k)$, bigrammes $p(W_k|W_{k-1})$, trigrammes $p(W_k|W_{k-1}, W_{k-2})$, et N-grammes $p(W_k|W_{k-1}, W_{k-2}, \ldots, W_{k-N})$ pour obtenir les meil-

[2]Personnellement, je m'en suis rendu compte dans les années 1985, après plusieurs mois durant lesquels le fichier contenant ces probabilités avait été lu à l'envers. La correction de cette erreur conduisit alors à une légère détérioration (non significative) des performance.

leurs estimateurs de N-grammes, avec une valeur de N pouvant aller jusqu'à 4 ou 5.

Bien que plusieurs équipes de recherche essaient toujours de s'attaquer aux problèmes de base et d'améliorer leur compréhension des problèmes sous-jacents à la reconnaissance automatique de la parole, il est cependant intéressant de s'arrêter un peu sur les thèmes de recherche actuellement les plus "populaires". Parmi ceux-ci, nous citerons notamment:

1. Accroissement de la taille des bases de données d'entraînement — Evidemment, l'augmentation de la taille des bases d'entraînement résulte (presque) toujours en une diminution du taux d'erreurs. Ceci est simplement du au fait que plus on a de données d'entraînement et plus on peut avoir des modèles détaillés (par la simple augmentation du nombre de paramètres). Dans ce cas, si la base de test n'est pas trop différente de la base d'entraînement, on observera effectivement une amélioration des performances. En conséquence, une des grandes tendances de ces 5 dernières années a été de générer des bases de données de plus en plus grandes, sans se poser trop de questions et en se limitant simplement aux contraintes purement logistiques (telles que taille mémoire et temps de calcul). Aujourd'hui, certaines bases de données sont suffisamment larges que pour estimer plus de 10^7 paramètres acoustiques et des "grammaires" de type trigrammes pour 64.000 mots. Malheureusement, très peu de scientifiques se sont arrêtés sur les problèmes de base. Quelles sont les meilleures données à utiliser? Quels sont les modèles les plus appropriés aux données? Comment réutiliser les données extraites de bases de données précédentes sur de nouveaux problèmes? Il serait plus intéressant d'avoir quelques réponses à ces questions plutôt que de simplement augmenter la taille des bases de données et de rapporter une réduction du taux d'erreurs.

2. Meilleur lissage des estimateurs statistiques — Evidemment, ce thème est étroitement lié au précédent. Avec suffisamment de données, des modèles très détaillés peuvent être entraînés sans avoir besoin de lisser les estimateurs de leurs nombreux paramètres. Cependant, avec les modèles actuels de reconnaissance de la parole, aucune base de données n'est jamais suffisamment large, et on utilise généralement plus de paramètres que nécessaire. Il faut donc toujours pratiquer le lissage des estimateurs, ce qui est généralement réalisé à l'aide de légères variantes de méthodes standards de lissage, conduisant à de légères variations dans l'amélioration des résultats.

3. Augmentation de la taille des vocabulaires — Malgré l'accroissement rapide des capacités de calcul et de mémoire des ordinateurs actuels, il reste toujours beaucoup de problèmes intéressants lorsque l'on essaye de reconnaître (en un temps "raisonable") des phrases prononcées à partir de grands vocabulaires. Malheureusement, jusqu'à présent, la plupart des travaux dans ce domaine se sont focalisés sur les problèmes purement techniques (temps de calcul, espace mémoire) plutôt que sur les véritables problèmes scientifiques tels que, par exemple, l'explosion du nombre de phrases acoustiquement similaires et l'accès lexical (au sens propre du terme, et pas simplement au sens purement informatique).

4. Accélération des techniques de décodage — Ceci est un exemple du type de travail requis par l'augmentation de la taille des lexiques. Ceci devient effectivement un travail important si on veut: (1) obtenir des démonstrations de systèmes de reconnaissance fonctionnant en temps réel, (2) pouvoir publier un "nombre" (en %) représentant le taux d'erreurs sur une base test de référence et (3) observer les erreurs de reconnaisance beaucoup plus vite. Tout cela, évidemment, en augmentant le nombre d'erreurs étant donné que l'accélération d'un système de reconnaissance se fait toujours au détriment de ses performances (déjà relativement faibles au départ).

5. Prononciations multiples — Beaucoup de travaux de recherche en cours essaient de développer des systèmes de reconnaissance capables d'utiliser les règles phonologiques et les différentes façons de prononcer les mots du vocabulaire. Il n'est cependant pas clair que les méthodes les plus sophistiquées conduisent aux meilleurs résultats. Bien souvent la conclusion est relativement simple: l'utilisation de quelques (entre 1 et 3?) variantes phonologiques (indépendantes) améliorent les taux de reconnaissance. Trop de prononciations possibles (ainsi que les représentations sous forme de graphes) diminuent généralement les performances. Une des raisons de cette observation est que l'augmentation du nombre de prononciations pour chaque mot du vocabulaire augmente auusi les risques de confusions entre mots ainsi que le nombre potentiels d'homonymes. Une deuxième raison possible est que le modèle utilisé (HMM), ainsi que la façon dont l'accès lexical est fait dans ce modèle, ne sont pas appropriés pour traiter ce type d'information.

6. Reconnaissance robuste de la parole — Bien que ce thème soit souvent présenté comme un sujet de recherche séparé, les recherches et dé-

veloppements dans ce domaine sont évidemment des plus importants pour un fonctionnement minimum dans des environnements réalistes. Tous les reconnaisseurs actuels sont réputés pour leur sensibilité extrême à différentes variables d'environnement (telles que bruits additifs et de convolution) ou d'élocution (telles que vitesse d'élocution et effort vocal), alors que l'oreille humaine est remarquablement robustes à celles-ci. Ces problèmes sont généralement abordés en améliorant la caractérisation du signal acoustique ou en adaptant les paramètres des modèles. Malheureusement, il ne semble pas que les méthodes développées jusqu'à présent soient vraiment satisfaisantes, et il n'est pas impossible que le modèle de base devra être modifié de façon significative de façon à pouvoir faire face aux différents types de variabilités qui n'ont pas été observés dans la base d'entraînement.

Malgré le caractère un peu critique de cette description, l'auteur tient à reconnaître ici que la technologie disponsible aujourd'hui peut déjà être suffisante pour développer avec succès quelques applications bien ciblés. Il est cependant également de notre opinion que tout nouveau progrès vraiment significatif dans ce domaine devra faire intervenir beaucoup plus que les thèmes abordés (de façon un peu caricaturale) ci-dessus.

Dans ce qui suit, nous discutons brièvement de quelques approches qui, nous le croyons, vont un peu plus dans le sens d'une meilleure modélisation du processus de reconnaissance. Cette présentation, fortement biaisée vers nos propres domaines d'activité, n'a cependant pas la prétention de fournir la solution ultime au problème mais simplement de présenter quelques voies de recherche qui pouraient peut-être nous permettre de sortir du minimum local dans lequel nous nous trouvons actuellement.

4. SYSTEME HYBRIDE HMM/ANN

Le but de ce papier n'est certainement pas de rediscuter des systèmes hybrides utilisant conjointement les modèles HMM et les réseaux de neurones (ANN). Cette technique est cependant la seule alternative actuellement capable de soutenir la comparaison avec les modèles HMM "classiques", tout en ayant l'avantage d'être plus simple (moins de modèles de phonèmes et moins de paramètres) et de permettre plus facilement certains types d'extensions.

Il a effectivement été montré (théoriquement et empiriquement) (Bourlard, 1990; Gish, 1990; Richard, 1991) que les ANNs pouvaient générer de bons estimateurs de probabilités à posteriori des classes de sorties (dans notre cas associées à des états de modèles HMMs) conditionnées sur les variables présentées à l'entrée du réseau. Cette propriété a permis à de nombreux chercheurs de développer un nouveau type de système de reconnaissance dans lequel les ANNs sont utilisés (par exemple, à la place de gaussiennes) pour générer les probabilités requises par les HMMs (Bourlard, 1994; Cohen, 1992; Lubensky, 1994; Robinson, 1991).

Dans ce papier, $X = \{x_1, \ldots, x_n, \ldots, x_N\}$ représente une séquence de N vecteurs acoustiques associés à une phrase particulière.

En quelques mots, l'approche hybride HMM/ANN peut se résumer comme suit. Comme dans les HMMs standards, chaque modèle est supposé construit à partir d'un ensemble de C classes $\Omega = \{\omega_1, \ldots, \omega_c, \ldots, \omega_C\}$ (avec lesquelles seront associées les densités de probabilité) qui seront utilisées pour construire une topologie (donnée a priori ou entraînée) de HMMs M_i pour chaque mot ou phrase à reconnaître. Chaque modèle HMM est donc défini comme un graph orienté contenant un certain nombre K d'états q^k $(k = 1, \ldots, K)$, chacun de ces états étant associé à une classe $\omega(q^k) \in \Omega$.

Dans le cas de HMMs standards, les fonctions de densité de probabilité requises $p(x_n|q^k) = p[x_n|\omega(q^k)]$ (aussi appelées fonctions de vraisemblance) sont typiquement supposées être des distributions gaussiennes ou multi-gaussiennes [3]. Dans le cas des systèmes hybrides HMM/ANN, un réseau de neurone contenant C unités de sortie (une pour chaque classe de Ω) est entraîné en mode de classification au niveau des vecteurs acoustiques pour générer des probabilités a posteriori locales $P[\omega_c|x_n, \Theta]$, $\forall c \in [1, C]$, où Θ représente l'ensemble des paramètres du réseau de neurone. Comme pour les HMMs standards, la segmentation de la base d'entraînement (en termes des classes de sortie) nécessaire à l'optimisation du réseau de neurones peut être obtenue de façon itérative par un algorithme de type Viterbi dont la convergence à été démontrée. Ces probabilités a posteriori locales sont ensuite utilisées, après division par les probabilités a priori $P[\omega_c]$ observées sur l'ensemble d'entraînement, comme des fonctions de vraisemblance $p[x_n|\omega(q^k), \Theta]/p(x_n)$ dans les HMMs pour calculer la vraisemblance globale $p(X|M, \Theta)$.

Ce type d'approche a été utilisé avec succès pour différents types de réseaux de neurones et différents types d'applications [voir, par exemple, (Bourlard, 1994; Cohen, 1992; Lubensky, 1994; Robinson, 1991)]. En dépit de sa simplicité, cette approche offre plusieurs avantages potentiels importants:

- Qualité du modèle: l'estimation des probabiltés par un réseau de neurones ne requiert aucune hypothèse concernant la forme des distributions statistiques sous-jacentes aux modèles. En théorie, les ANNs peuvent en effet générer n'importe quelle fonction non linéaire.

[3] Dans ce papier, les probabilités seront dénotées $P(.)$ tandis que les fonctions de densité de probabilité seront dénotées $p(.)$.

- Discrimination: les réseaux de neurones sont généralement entraînés selon un critère discriminant, ce qui améliore les performances du système.

- Information contextuelle et corrélation: les réseaux de neurones peuvent facilement tenir compte de plusieurs vecteurs acoustiques successifs et en modéliser la corrélation (temporelle) éventuelle. Si X_{n-d}^{n+e} représente une sous-séquence $\{x_{n-d}, \ldots, x_n, \ldots, x_{n+e}\}$, les réseaux de neurones peuvent estimer des probabilités de type $P[\omega_c|X_{n-d}^{n+e}]$. Pour différentes raisons, cela n'était pas possible avec les HMMs standards, les solutions les plus proches étant: (1) l'utilisation des dérivées temporelles et (2) l'utilisation de quelques vecteurs adjacents sur lesquels on applique une analyse discriminante linéaire pour réduire la dimension de l'espace et en extraire les caractéristiques essentielles avant leur utilisation dans les HMMs (Hunt, 1989; Haeb-Umbach, 1994).

- Meilleure utilisation des paramètres: en effet, toutes les classes partagent le même ensemble de paramètres pour générer les fonctions discriminantes non linéaires optimales pour la classification. Il est également connu en statistique qu'il est plus économique de modéliser les frontières de classes (ce que fait les ANNs) plutôt que les fonctions de densités.

- Meilleure flexibilité: l'utilisation d'un réseau de neurones comme estimateur de probabilités permet de combiner aisément différentes variables et caractéristiques du signal.

- Complémentarité: les réseaux de neurones peuvent parfois fournir une information complémentaire à celle obtenue à la sortie d'un système statistique (HMM) classique. Ceci est le cas, par exemple, lors de la combinaison des HMMs avec un réseaux de neurones particulier appelé "segmental neural network" (Austin, 1992). Dans ce cas, les HMMs sont utilisés pour générer les N meilleures phrases dont les segmentations sous-jacentes sont ensuite réestimées en utilisant un réseau de neurones particulier qui prend en compte des segments phonétiques entiers (permettant ainsi de tenir
  compte des corrélations temporelles internes). Il a été montré que ce type d'approche pouvait améliorer les taux de reconnaissance.

Beaucoup de systèmes de reconnaissance hybrides relativement simples ont été récemment développés. Sur différents tests de référence, ces systèmes ont démontré leurs qualités au niveau des performances (comparables aux meilleurs systèmes HMM), ainsi que du temps de calcul et de l'espace mémoire requis pour la reconnaissance [voir, par exemple, (Lubensky, 1994; Robinson, 1993)].

Plus récemment, un tel système [ABBOT de l'Université de Cambridge (Hochberg, 1995)] a été évalué dans le cadre du programme ARPA aux USA et du projet (LRE) européen SQALE (20.000 mots de vocabulaire, indépendance au locuteur, parole continue). Les résultats de l'évaluation (Steeneken, 1995) affichent un léger avantage du système hybride par rapport aux autres systèmes européens, tout en nécessitant beaucoup moins de CPU pour effectuer les tests de reconnaissance. Un autre résultat intéressant est que les modèles acoustiques de ce système n'utilisent que quelques centaines de milliers de paramètres (aux alentours de 500.000 pour ABBOT), alors que les autres systèmes plus classiques mettent en jeu plusieurs millions (environ 10^7) de paramètres.

Le seul point négatif de cette approche hybride est le temps de calcul requis pour l'entraînement des modèles ANN. Ceci a certainement été un handicap au développement initial de ces systèmes (pour lesquels du matériel spécial a du être développé). Il est cependant bon de noter que, étant donné l'évolution rapide des moyens de calculs, ceci est de moins en moins une limitation. Les réseaux de neurones typiquement utilisés pour les grandes bases de données peuvent maintenant être entraînés sur ces mêmes bases de données en quelques jours sur des stations de travail modernes.

5. MEILLEURE MODELISATION?

5.1. Accepteur Stochastique

De récents développements suggèrent une autre alternative aux systèmes HMMs, permettant une généralisation intéressante des systèmes hybrides HMM/ANN. Pour ne pas ennuyer le lecteur, nous ne donnerons qu'une description très succinte de la théorie du modèle (ceci n'étant en effet pas le but de ce papier) avant d'en discuter les avantages.

Cette nouvelle approche, plutôt que d'utiliser un modèle de production tel que le HMM, avec émission de vecteurs acoustiques sur les états, est basée sur la notion d'accepteur stochastique à nombre d'états fini (SFSA, pour "Stochastic Finite State Acceptor"). Dans ce cas, la transition (stochastique) d'un état vers un autre de l'automate correspond à la reconnaissance (l'"acceptation") d'un vecteur acoustique. Ces probabilités, comme initialement définies dans (Bourlard, 1994), découlent directement du développement (sous certaines hypothèses) de $P(M|X)$ qui peut s'exprimer en fonction de *probabilités de transition conditionnelles*:

$$P(q_n^t|q_{n-1}^k, x_n) = P(\omega(q^t)|\omega(q^k), x_n)$$

où q_n^k représente l'événement d'accepter l'état q^k à l'instant n, $\omega(q^k)$ étant la classe (distribution statis-

tique) associée à cet état. Ces probabilités peuvent être estimées grâce à un réseau de neurones particulier dans lequel l'entrée à été étendue pour tenir compte de la classe précédente.

L'entraînement de ces probabilités de transition conditionnelles peut être réalisé selon une procédure baptisée REMAP (Recursive Estimation and Maximization of A Posteriori probabilities), une approche itérative qui garantit l'accroissement des probabilités a posteriori des modèles corrects associés aux phrases d'entraînement (Bourlard, 1995). Ceci résulte donc en un système plus discriminant. En effet, la somme de toutes les probabilités a posteriori (sur tous les modèles possibles) étant égale à 1, l'augmentation de la probabilité du modèle correct résulte nécessairement en une diminution des probabilités des modèles rivaux.

En quelques mots, l'entraînement REMAP consiste à trouver les paramètres Θ du réseau de neurones maximisant:

$$\prod_{j=1}^{J} P(M_j|X_j,\Theta) \qquad (1)$$

pour toutes les phrases d'entraînement X_j et leurs modèles associés M_j $(j = 1,\ldots,J)$. Pour résoudre ce problème, il faut donc se poser la question suivante: étant donné un réseau de neurones de paramètres Θ estimant $P(\omega_\ell|x_n,\omega_k,\Theta)$, quelles doivent être les nouvelles sorties cibles à utiliser pour nous garantir qu'un nouvel entraînement de ce réseau nous générera de nouveaux estimateurs $P(\omega_\ell|x_n,\omega_k,\Theta^*)$ conduisant à une augmentation des probabilités a posteriori globales pour les bons modèles. Dans (Bourlard, 1995), nous montrons que la nouvelle cible pour la sortie associée à ω_ℓ à l'instant n (pour le vecteur acoustique x_n) et l'état précédent q^k est donnée par

$$P(\omega(q_n^\ell)|X,\omega(q_{n-1}^k),\Theta,M) \qquad (2)$$

correspondant au meilleur estimateur local (pour pouvoir entraîner le réseau de neurones), tenant compte de la phrase complète X. Il est possible de montrer qu'une réestimation itérative de ces nouvelles cibles (et nouveaux paramètres Θ) converge vers un maximum (local) des probabilités a posteriori globales. Dans (Bourlard, 1995), une méthode rapide pour calculer $P(\omega(q_n^\ell)|X,\omega(q_{n-1}^k),M)$ en fonction de $P(\omega(q^\ell)|x_n,\omega(q^k))$ est présentée.

En plus des meilleures propriétés de discrimination discutées plus haut, ce modèle SFSA/REMAP a plusieurs autres avantages potentiels:

- Meilleures propriétés de modélisation (plutôt que la simple description) — En effet, contrairement aux HMMs standards où la modélisation de toute nouvelle variable (par exemple, la dépendance au contexte phonétique) résulte généralement en une augmentation du nombre de modèles et du nombre de paramètres. Dans le cas des SFSAs, les observations acoustiques, aussi bien que toute autre variable pouvant affecter ces observations (en d'autres mots toutes les variables pouvant affecter la réalisation phonétique ou autre unité élémentaire de parole), apparaissent dans la conditionnelle des probabilités de transition. Etant donné cette propriété, on peut espérer que les facteurs influençant les réalisations phonétiques (par exemple, le contexte phonétique, la vitesse d'élocution, l'effort voval, etc...) pourraient être modélisés sans explosion du nombre de paramètres. Evidemment, ceci pourrait, en principe, se faire avec n'importe quel estimateur statistique. Il est cependant probable que la réduction du nombre de paramètres (modèles) se fera au prix d'une plus grande complexité (non linéaire) des relations statistiques. Comme discuté dans la section précédente, les réseaux de neurones se sont cependant montrés particulièrement efficaces pour l'estimation de distributions de probabilités complexes avec un nombre limité de paramètres.

- Le formalisme utilisé pour les SFSAs, ainsi que leur algorithme d'entraînement discriminant, conduit à un modèle dans lequel les paramètres des modèles acoustiques (probabilités de transition conditionnelles) et certaines informations du modèle de language sont intimement liés. Par modèle de language nous entendons ici toute information relative à la topologie des modèles, incluant notamment la structure phonétique, la structure phonologique (mots en termes de phonèmes), et la structure syntaxique (phrases en termes de mots). Alors que ces informations sont généralement extraites indépendamment des modèles acoustiques (via des connaissances explicites ou de grandes bases de données de texte), les systèmes SFSAs/REMAP apprennent implicitement ces informations telles qu'elles sont présentes dans la base d'entraînement acoustique. Il n'est cependant pas certain que cela soit toujours bénéfique.

5.2 Reconnaissance Multi-Bande

Les systèmes standards de reconnaissance de la parole traitent chaque vecteur acoustique (par exemple, un vecteur spectral calculé sur un segment de 30 ms, toutes les 10 ms) comme une entité. En conséquence, même si une seule bande de fréquence est bruitée, tout le vecteur acoustique est contaminé et, typiquement, les performances du reconnaisseur en sont fortement affectées (même si le bruit n'affecte pas les bandes de fréquence les plus importantes pour la reconnaissance).

En plus de cette observation, les travaux de Fletcher et ses collègues (Fletcher, 1953), récemment rappelés dans (Allen, 1994), suggèrent que le processus de reconnaisance humain est basé sur des sous-bandes de fréquences qui sont traitées indépendamment les unes des autres. Les décisions partielles (en fréquence) sont alors recombinées à "un certain niveau temporel" (au niveau de certaines unités acoustiques) de façon à ce que le taux d'erreurs global (tenant compte de toute la bande de fréquence) soit égal au produit des différents taux d'erreurs partiels (pour chaque sous-bande de fréquence). Bien que facile à formuler, cette règle est cependant impossible à mettre en pratique de façon stricte. En effet, elle nécessiterait la connaissance formelle de la meilleure bande de fréquence.

Il est également connu que l'oreille humaine est relativement indépendante de la phase (jusqu'à un maximum de l'ordre de 200 ms), ce qui n'est pas le cas de nos reconnaisseurs de parole.

En plus de ces considérations psycho-acoustiques (qui sont parfois remises en question), nous voyons plusieurs raisons techniques pour considérer cette approche en sous-bandes:

1. Meilleure robustesse dans le cas de bruits non observés sur les bases d'entraînement mais n'affectant que quelques bandes de fréquence spécifiques.

2. Certaines sous-bandes de fréquence sont plus critiques que d'autres pour la reconnaissance de certains sons.

3. En regardant les sonogrammes, nous voyons que les transitions (entre phonèmes?) ne se produisent pas toujours au même instant pour toutes les bandes de fréquence. Cette "pente spectrale" est bien connue (et varie notamment en fonction de l'effort vocal) et reflète une asynchronie entre les différentes bandes de fréquence. Les avantages potentiels de l'approche en sous-bandes sont: (1) l'insensibilité (dépendant du niveau de recombinaison) à cette asynchronie, ainsi qu'à la phase en général et (2) une plus grande "stationarité" des signaux à bande de fréquence limitée (d'où une modélisation HMM plus simple).

4. Des stratégies différentes de reconnaissance pourraient être appliquées à différentes bandes de fréquence (par exemple, différentes caractéristiques, différentes constantes de temps, ...).

Evidemment, pour éviter de permettre trop de flexibilité dans les décisions en sous-bandes (dans chacune desquelles il y a évidemment moins d'information que dans la bande de fréquence complète), il est alors nécessaire de réinjecter des contraintes à un plus haut niveau. Dans notre cas, cette contrainte est introduite en forçant la synchonie (en termes de segmentation sous-jacente) des différentes bandes de fréquence à un certain niveau. Les différents niveaux possibles sont 'l' état HMM, le phonème ou le mot, bien que certaines considérations semblent indiquer que la syllabe pourrait être un bon candidat pour ce type d'approche.

Bien que le critère de recombinaison à la Fletcher suggère une règle de recombinaison optimale, nous ne connaissons aucune méthode statistique capable de modéliser cette règle. Dans notre cas, nous avons donc étudié la recombinaison des bandes de fréquence selon la loi:

$$P(X|M) = f(\{w_k\}, \{P(X_k|M_k)\}) \qquad (3)$$

où M représente le modèle HMM (ou HMM/ANN) de mot ou de phrase, M_k ce même modèle de mot ou de phrase associé à la k-ième bande de fréquence, et X_k la séquence de vecteurs acoustiques (bandes critiques) réduite à cette k-ième bande de fréquence. Deux fonctions de recombinaison $f(\cdot)$ ont été testées:

$$f(\cdot) = \sum_{k=1}^{K} w_k \log P(X_k|M_k) \qquad (4)$$

dans laquelle les w_k sont les paramètres de recombinaison linéaire, et

$$f(\cdot) = MLP_{w_k}(\log P(X_k|M_k)) \qquad (5)$$

où MLP_{w_k} représente alors un perceptron multi-couche (MLP, pour multilayer perceptron) utilisé pour la recombinaison et paramétrisé en terme des w_k's et ayant les $\log P(X_k|M_k)$, $\forall k$, à son entrée.

Dans l'expérience brièvement discutée ci-dessous, un système hybride HMM/ANN de phonèmes à 3 états et 18 bandes critiques a été utilisé comme système de référence. Pour les reconnaisseurs en sous-bandes, 3 sous-bandes (à partir des 18 bandes critiques) avec recouvrement partiel, respectivement de [0-1200], [660-2490], et [1480-4000] Hz ont été utilisées. Chaque sous-bande a été traitée par un reconnaisseur
HMM/ANN phonétique. La base de données utilisée ici a été enregistrée sur ligne téléphonique à partir d'un vocabulaire de 108 mots isolés (en allemand). Un ensemble de 15 locuteurs prononçant une fois chaque mot a été utilisé pour les tests.

Deux niveaux de recombinaison ont été testés: recombinaison au niveau de l'état HMM et recombinaison au niveau du mot. Pour la recombinaison au niveau de l'état HMM, trois différentes approches ont été comparées sur base de (4):

1. Pas de pondération ("No-W" dans la Table 1): tous les poids w_k dans (4) sont mis à 1.

2. Poids w_k dans (4) proportionnels aux taux de reconnaissance phonétique (au niveau du vect-

	FB	No-W	Acc-W	SNR-W	MLP
clean	3.6%	3.7%	3.7%	3.2%	2.7%
noisy	25.5%	9.2%	6.7%	6.3%	—

Table 1: *Reconnaissance mots isolés (108 mots en allemand, ligne téléphonique). "FB" correspond au reconnaisseur standard (full band); "No-W" correspond au reconnaisseur en sous-bandes avec recombinaison au niveau de l'état sans pondération; "Acc-W"= recombinaison au niveau de l'état avec les poids proportionnels au taux de reconnaisance phonétique dans chaque sous-bande; "SNR-W"= recombinaison au niveau de l'état avec des poids inversement proportionnels au rapport signal/bruit dans chaque sous-bande. La colonne "MLP" se rapporte à la recombinaison au niveau du mot, utilisant un MLP optimisé sur la base d'entraînement.*

eur acoustique) dans chaque sous-bande de fréquence, et calculé uniquement sur l'ensemble d'entraînement ("Acc-W" dans la Table 1).

3. Poids w_k dans (4) inversement proportionnels au rapport signal sur bruit (dans chaque sous-bande) estimé automatiquement par une technique "on-line" sur l'ensemble test ("SNR-W" dans la Table 1).

Pour la recombinaison au niveau du mot, un MLP a été utilisé selon (5) ("MLP" dans la Table 1). Dans ce cas, le MLP avait donc 108 (mots) $\times$ 3 (bandes) à son entrée et 108 sorties.

Après avoir entraîné les différents systèmes sur de la parole non bruitée (telle qu'enregistrée sur les lignes téléphoniques), les différents reconnaisseurs ont été testés sur la parole non bruitée ainsi que sur les mêmes données contaminées par du bruit additif sélectif (bruit blanc dans la 1ère sous-bande, 10dB de rapport signal/bruit). Les énergies des bandes critiques ont été utilisées comme caractéristiques du signal.

Considérant les résultats donnés dans la Table 1, il semble que pour le signal non bruité il soit possible d'obtenir des résultats équivalents ou meilleurs (MLP) que le système classique en décomposant le problème en sous-bandes. Ceci est déjà intéressant en soi étant donné que ca montre que le système standard, basé sur des bandes de fréquence synchrones, n'est pas nécessairement optimal. Lorsqu'une des bandes de fréquence est contaminée par du bruit additif, l'approche multi-bande semble alors être beaucoup plus robuste. Dans le cas de la recombinaison linéaire au niveau de l'état, les meilleurs résultats sont obtenus en utilisant les poids proportionnels au rapport signal/bruit dans chaque sous-bande. Il est cependant remarquable que, même sans ces poids, et en utilisant uniquement des poids obtenus sur l'ensemble d'entraînement (ou même pas de poids de tout!), les performances du système restent très bonnes. Même comparée avec des techniques de soustraction de bruit (résultats non rapportés ici pour ne pas ennuyer le lecteur), l'approche en sous-bandes reste supérieure pour la base de données étudiée ici.

Etant donné que rien n'a encore été optimisé dans cette approche (nombre de sous-bandes, niveau de recombinaison, critère de recombinaison), il y a donc lieu de croire que ce système pourrait avoir quelques potentialités.

6. DISCUSSION ET CONCLUSION

Malgré certaines apparences, le but de ce papier n'était pas de présenter des résultats techniques, mais simplement de prêcher en faveur d'un peu plus d'originalité, et de risque (!), dans la recherche en reconnaissance de la parole. Nous craignons effectivement que les recherches actuelles se focalisent trop sur l'approche dominante des HMMs. Bien qu'étant une très bonne technique de base, beaucoup d'entre nous ont cependant l'impression que celle-ci a atteint sa maturité et qu'il sera nécessaire d'aller au-delà de l'augmentation du nombre de paramètres et de la taille des bases d'entraînement pour véritablement résoudre le problème de la reconnaissance robuste de la parole. Tout en conservant l'acquis énorme accumulé pendant ces 10 dernières, nous croyons qu'il serait bon maintenant de se focaliser sur les points faibles des systèmes existants. Pour dépasser ceux-ci, il sera sans doute nécessaire d'abandonner le confort d'une technique bien maîtrisée.

De façon certainement très biaisée, et simplement comme exemples, certaines des approches sur lesquelles nous travaillons actuellement ont été brièvement présentées. Evidemment, beaucoup de celles-ci n'aboutiront peut-être jamais à des résultats concluants. Dans ce cas, nous espérons que nous aurons au moins appris quelque chose en chemin. Nous savons que beaucoup de collègues travaillent également sur d'autres approches originales. Tout comme nous, ils ont sans doute difficile de faire mieux que les systèmes standards qui ont été optimisés pendant 10-20 ans. Il ne faut cependant pas abondonner; il faut en effet beaucoup de temps et beaucoup de persévérance pour arriver à des résultats, soit bons, soit instructifs.

7. REMERCIEMENTS

La plupart des travaux discutés ici sont le résultat d'une collaboration très étroite avec le groupe de Nelson Morgan (Intl. Computer Science, Berkeley, CA) et de Hynek Hermansky (Oregon Graduate Institute, Portland, OR). Les résultats expérimentaux présentés à la Section 5.2 on été obtenus par Stéphane Dupont de la Faculté Polytechnique de Mons. Certaines vues présentées dans ce papier ont bénéficié

de discussions avec de nombreux collègues. Qu'ils en soient tous remerciés.

8. BIBLIOGRAPHIE

Allen, J.B. (1994). "How do humans process and recognize speech?," *IEEE Trans. on Speech and Audio Processing*, vol. 2, no. 4, pp. 567-577.

Austin, S., Zavaliagkos, G., Makhoul, J., and Schwartz, J. (1992). "Improving State-of-the-Art Continuous Speech Recognition Systems Using the N-best Paradigm with Neural Networks," Proc. DARPA Speech and Natural Language Workshop, Harriman, NY, Morgan Kaufmann, pp. 180-184.

Baker, J.K. (1975) "The Dragon system – an overview," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. ASSP-23, no. 1, pp.24-29.

Bourlard, H. and 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.

Bourlard, H. and Morgan, N. (1994). *Connectionist Speech Recognition — A Hybrid Approach*, Kluwer Academic Publishers.

Bourlard, H., Konig, Y., and Morgan, N. (1995). "REMAP: recursive estimation and maximization of a posteriori probabilities in connectionist speech recognition," *Proc. of Eurospeech'95* (Madrid, Spain).

Cohen, M., Franco, H., Morgan, N., Rumelhart, D. and Abrash, V. (1992). "Hybrid neural network/hidden Markov model continuous speech recognition," *Proc. of Intl. Conf. on Speech and Language Processing* (Banff, Canada), pp. 915-918.

Fletcher, H. (1953). *Speech and Hearing in Communication*, New York: Krieger.

Gish, H. (1990). "A probabilistic approach to the understanding and training of neural network classifiers," *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing* (Albuquerque, NM), pp. 1361-1364.

Haeb-Umbach, R., Geller, D., Ney, H. (1994). "Improvements in connected digit recognition using linear discriminant analysis and mixture densities," *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing* (Adelaide, Australia), pp. II-239-242.

Hochberg, M.M., Renals, S.J., Robinson, A.J., and G.D. Cook. (1995). "Recent Improvements to the ABBOT Large Vocabulary CSR System," *Proc. of IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing* (Detroit, MI), pp. 69-72.

Hunt, M. and Lefebvre, C. (1989). "A comparison of several acoustic representations for speech recognition with degraded and undegraded speech," *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing* (Glasgow, Scotland), pp. 262-265.

Jelinek, F., Bahl, L.R., and Mercer, R.L. (1975). "Design of a linguistic statistical decoder for the recognition of continuous speech," *IEEE Trans. Information Theory*, vol. IT-21, pp. 250-256.

Lubensky, D.M., Asadi, A.O., and Naik, J.M. (1994). "Connected digit recognition using connectionist probability estimators and mixture-Gaussian densities," *Proc. of the Intl. Conf. on Spoken Language Processing* (Yokohama, Japan).

Richard, M.D. and Lippmann, R.P. (1991). "Neural network classifiers estimate Bayesian a posteriori probabilities," *Neural Computation*, no. 3, pp. 461-483.

Robinson, T. and Fallside, F. (1991). "A recurrent error propagation network speech recognition system," *Computer Speech and Language*, no. 5, pp. 259–274.

Robinson, T., Almeida, L., Boite, J.M., Bourlard, H., Fallside, F., Hochberg, M., Kershaw, D., Kohn, P., Konig, Y., Morgan, N., Neto, J.P., Renals, S., Saerens, M., and Wooters, C. (1993). "A neural network based, speaker independent, large vocabulary, continuous speech recognition system: the Wernicke project," *Proc. EUROSPEECH'93* (Berlin, Germany), pp. 1941-1944.

Steeneken, J.M. and Van Leeuwen, D.A. (1995). "Multi-lingual assessment of speaker independent large vocabulary speech recognition systems: the *SQALE* project (speech recognition quality assessment for language engineering)," *Proceedings of EUROSPEECH'95* (Madrid, Spain), pp. 1271-1274.

AMELIORATION DES PERFORMANCES DE REJET PAR APPRENTISSAGE DISCRIMINANT

Hugues Leprieur

France Télécom CNET-LAA/TSS/RCP

Technopole Anticipa, 2 avenue Pierre Marzin, 22307 LANNION Cedex*

Tél.: 96 05 38 70 - Fax 96 05 35 30

ABSTRACT

Out-of-vocabulary utterances modeling remains a strong weakness in HMM-based automatic speech recognition systems. Such a modeling may be improved by discriminative training. This paper presents a comparison between two discriminative criteria optimized by minimum memory loss optimisation technique, which provides up to a 37 % error rate reduction.

1. INTRODUCTION

La qualité d'un serveur vocal commandé par la voix dépend largement de sa capacité à rejeter la parole n'appartenant pas au vocabulaire de l'application, ainsi que les bruits.

Dans de nombreux systèmes de reconnaissance à base de modèles de Markov cachés (HMM), la technique de rejet utilisée consiste à modéliser la parole n'appartenant pas au vocabulaire de l'application par des modèles poubelles. Cependant, l'optimalité de l'apprentissage classique au maximum de vraisemblance repose sur un modèle correct de génération des données, hypothèse d'autant moins réaliste que l'on n'utilise que quelques modèles grossiers pour modéliser tous les événements acoustiques autres que les mots clés.

Les techniques d'apprentissage discriminant par descente de gradient peuvent résoudre ce problème; notamment, elles sont capables de traiter des hypothèses incorrectes de modélisation (Brown, 87).

Récemment, ces techniques ce sont révélées efficaces lorsqu'elles étaient utilisées pour corriger des modèles estimés au préalable selon le critère du maximum de vraisemblance (MLE).

En particulier pour une application téléphonique de reconnaissance des chiffres espagnols, dans (Torre, 94), on affinait les paramètres de manière à minimiser une fonction de coût constituée de la somme pondérée des estimées des probabilités de fausses alarmes, d'omissions et de substitutions. Une technique similaire a été utilisée par (Rose, 95) pour affiner un système d'acceptation/rejet des hypothèses de détection sur la base de données *Switchboard.*

2. APPRENTISSAGE DISCRIMINANT

Alors que les algorithmes de Baum-Welch et de Viterbi pour l'estimation selon le critère du maximum de vraisemblance visent à maximiser indépendamment pour chaque modèle la probabilité d'émission des observations qu'il génère, l'apprentissage discriminant vise à maximiser explicitement le taux de reconnaissance, en essayant d'augmenter l'écart entre la probabilité d'émission du modèle correct et celle des modèles compétiteurs. Ce paragraphe présente les deux critères que nous essayons de maximiser par apprentissage discriminant en modifiant les paramètres Θ des HMM. Les performances obtenues après optimisation de l'un ou l'autre des critères sont fournies au paragraphe 4.

2.1. Maximiser l'information mutuelle

Minimiser pour la classe correcte, l'entropie croisée entre la probabilité *a posteriori* estimée et celle désirée revient à maximiser l'information mutuelle entre l'observation acoustique et le modèle (Ney, 95).

$$F_{MMI}(\Theta) = \sum_{n=1}^{N} \ln P_{\Theta}(K_n / X_n)$$

où K_n représente le modèle correct (c'est-à-dire le modèle du mot contenu dans l'observation acoustique X_n), N est le nombre total d'exemples dans la base d'apprentissage, et $P_{\Theta}(.)$ représente la probabilité estimée par le système.

2.2. Minimiser l'erreur de classification

L'objectif de tout système de reconnaissance étant de minimiser la

* adresse actuelle : A.T.N.E., 7 av. de l'Atlantique, 91940 Les Ulis

probabilité d'erreur, il apparaît intéressant d'optimiser directement les paramètres du modèle sur ce critère.

Nous sommes cependant confrontés à deux problèmes. Le premier est que l'erreur est binaire en ce sens que soit il y a erreur soit il n'y a pas erreur. Or il nous faut trouver un critère dérivable si l'on veut modifier les paramètres par descente de gradient.

Nous souhaitons également que notre système ait une capacité à généraliser, c'est-à-dire qu'il classe correctement des données qui ne font pas partie de l'ensemble d'apprentissage.

Optimiser le critère du minimum d'erreur de classification (MCE) constitue une réponse au problème (Ney, 95):

$$F_{MCE}(\Theta) = \sum_{n=1}^{N} \frac{1}{1+\left[\frac{Q_\Theta(K_n,X_n)}{P_\Theta(K_n,X_n)}\right]^s}$$

avec:

$$Q_\Theta(K_n,X_n) = \left[\sum_{M\neq K_n} [P_\Theta(M,X_n)]^r\right]^{1/r}$$

où $r \geq 1$ et $s \geq 0$.

Dans les expériences nous avons fixé $r = 1$ et $s = 1$.

2.3. Technique d'optimisation

La technique d'optimisation que nous avons choisie consiste à partir d'un HMM dont les paramètres (probabilités de transition et paramètres des fonctions de densité de probabilité d'observation gaussiennes) ont été estimés selon le critère du maximum de vraisemblance. Ensuite, on modifie les moyennes et écarts-types des gaussiennes de manière à optimiser le critère MMI ou MCE.

Bien que cette optimisation puisse se faire par descente de gradient, nous avons constaté de meilleurs résultats en utilisant la technique du minimum d'oubli (MML) (Leprieur, 95).

La modification des paramètres du système se fait une seule fois par époque, c'est-à-dire que l'alignement des enregistrements n'est remis en cause qu'après présentation de toute la base de données. Le pas d'apprentissage reste constant entre deux époques.

3. PROTOCOLE EXPERIMENTAL

3.1. Apprentissage de base

Les HMM utilisés dans les deux expériences sont constitués de 4 modèles poubelles et d'un réseau allophonique décrivant les mots clés (Jouvet, 91). Le langage de l'application est décrit par une grammaire nulle qui ne permet au système que de reconnaître un seul mot clé.

Les modèles poubelles sont des modèles de Bakis à 30 états et 30 monogaussiennes. Les trames acoustiques sont calculées toutes les 16ms, et sont constituées de 8 coefficients MFCC et de l'énergie, ainsi que de leurs dérivées première et seconde. Les modèles sont entraînés par l'algorithme de Viterbi à la fois sur les données de laboratoire et d'exploitation. Les modèles des mots clés sont appris sur les enregistrements contenant une prononciation correcte du mot clé, alors que les modèles poubelles sont appris sur des données hors-vocabulaires ou bruits.

3.2. Bases de données

Les expériences décrites dans cet article ont été réalisées sur deux bases de données distinctes (cf. tables 1 et 2), qui sont constituées des enregistrements d'appels aux serveurs vocaux en exploitation *Les Baladins* et *Nîmes*, capables de reconnaître respectivement 26 et 9 mots clés (Athimon, 94). Afin que chaque mot soit suffisamment représenté dans le corpus d'apprentissage, en plus des enregistrements d'*exploitation* nous avons ajouté des enregistrements de *laboratoire*.

Table 1 : Base de données *Les Baladdins*

	Nb d'enregistrements	
	Mots-clés	Hors vocab.
Ensemble d'apprentissage	12013 exp 9918 lab.	3279 exp
Ensemble de test (exploitation)	12083	3309

Table 2 : Base de données *Nîmes*

	Nb d'enregistrements	
	Mots-clés	Hors vocab.
Ensemble d'apprentissage	3658 exp 2695 lab.	7451 exp
Ensemble de test (exploitation)	3598	8242

3.3. Mesure des taux d'erreur

La mesure des performances se fait par le calcul des taux d'erreur suivants : le taux de substitution entre mot clés (sub), le taux d'omission (omi), le taux de fausses

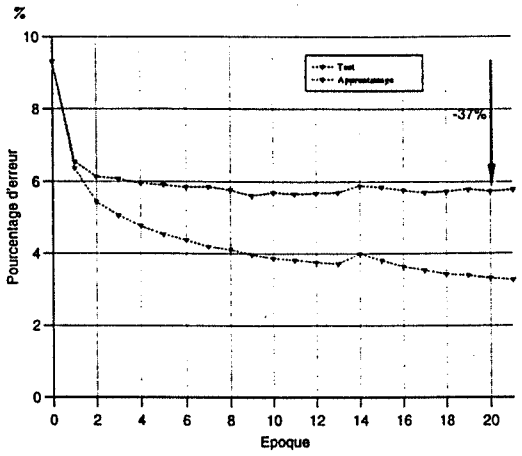


Figure 2: Réapprentissage MCE

La figure 2 montre l'évolution du taux d'erreur global en fonction des époques d'apprentissage sur le corpus utilisé pour l'apprentissage et le test. La réduction de 37% du taux d'erreur nous fournit une borne supérieure de la réduction que l'on peut espérer réaliser après mise en place d'une procédure de validation croisée.

Un des arguments souvent présentés à l'encontre des techniques d'apprentissage discriminant est que par rapport au MLE, elles présentent davantage de risques de spécialiser les modèles aux bases de données. Afin de vérifier cette hypothèse, nous avons fourni à l'entrée du même système (MCE sur la base de *Nîmes*) une autre base de données appelée *Trégor* qui ne contient que des mots à rejeter.

La base *Trégor* contient 34 mots mots n'appartenant pas au vocabulaire de *Nîmes* et qui ont été prononcés par 365 locuteurs.

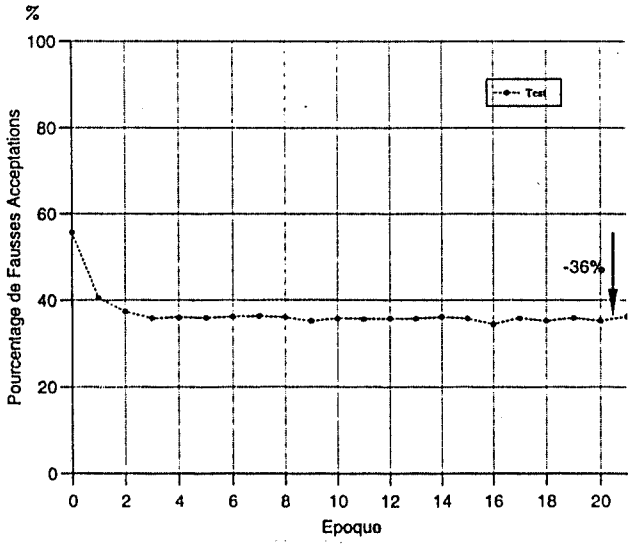


Figure 3: Rejet de la base du *Trégor*

La figure 3 montre que l'on obtient là encore une amélioration significative du taux de rejet.

5. CONCLUSION

Les résultats montrent qu'un réapprentissage discriminant afin de maximiser les critères MCE et MMI permet d'obtenir une amélioration significative des performances sur deux applications utilisant une modélisation grossière de la parole à rejeter, semblable à celle qui est utilisée dans deux serveurs vocaux en exploitation.

Un avantage supplémentaire d'une telle méthode est qu'elle n'engendre aucun surcoût durant la phase de reconnaissance.

6. BIBLIOGRAPHIE

Rose R.C., Juang B.H, et Lee C.H. (1995), A training procedure for verifying string hypothesis in continuous speech recognition, *IEEE International Conference on ASSP, Detroit, mai 1995, vol.I, pp. 281-284.*

De la Torre C. et Acero A. (1994), Discriminative training of garbage model for non-vocabulary utterance rejection, *International Conference on Spoken Language Processing, Yokohama, Japon, septembre 1994, vol.I, pp. 475-478.*

Athimon C., Bigorgne D., Cherbonnel B., Dubois D., Gagnoulet C., Jouvet D., Marzio H., Monné J., Py S., Sorin C. et. Toularhoat M. (1994), Operational and experimental french telecommunication services using CNET speech recognition and text-to-speech synthesis, *IEEE Workshop on Interactive Voice Technology for Telecommunications Applications, Kyoto, Japon, septembre 1994, pp. 27--32.*

Ney H. (1995), On the probabilistic interpretation of neural network classifiers and discriminative training criteria, *IEEE Transactions on PAMI, février 1995, vol.17, n2 pp. 107-119.*

Brown P.F. (1987), The Acoustic-Modelling Problem in Speech Recognition, *PhD thesis, Carnegie Mellon University, Pittsburgh, USA.*

.Jouvet D., Bartkova K. et Monné J. (1991) On the modelization of allophones in an HMM based speech recognition system, *2nd European Conference on Speech Communication and Technology,* Gênes, *Italie, vol. 2, pp. 923-926.*

Leprieur H. et P. Haffner (1995), Discriminant learning with minimum memory loss for improved non-vocabulary rejection, *4th European Conference on Speech Communication and Technology, Madrid, Espagne, vol.1, pp. 89-92.*

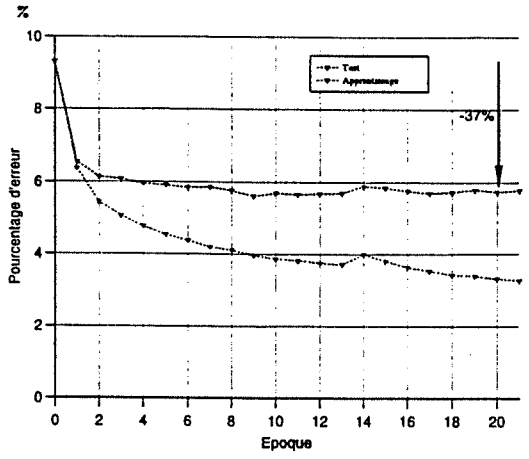


Figure 2: Réapprentissage MCE

La figure 2 montre l'évolution du taux d'erreur global en fonction des époques d'apprentissage sur le corpus utilisé pour l'apprentissage et le test. La réduction de 37% du taux d'erreur nous fournit une borne supérieure de la réduction que l'on peut espérer réaliser après mise en place d'une procédure de validation croisée.

Un des arguments souvent présentés à l'encontre des techniques d'apprentissage discriminant est que par rapport au MLE, elles présentent davantage de risques de spécialiser les modèles aux bases de données. Afin de vérifier cette hypothèse, nous avons fourni à l'entrée du même système (MCE sur la base de *Nîmes*) une autre base de données appelée *Trégor* qui ne contient que des mots à rejeter.

La base *Trégor* contient 34 mots mots n'appartenant pas au vocabulaire de *Nîmes* et qui ont été prononcés par 365 locuteurs.

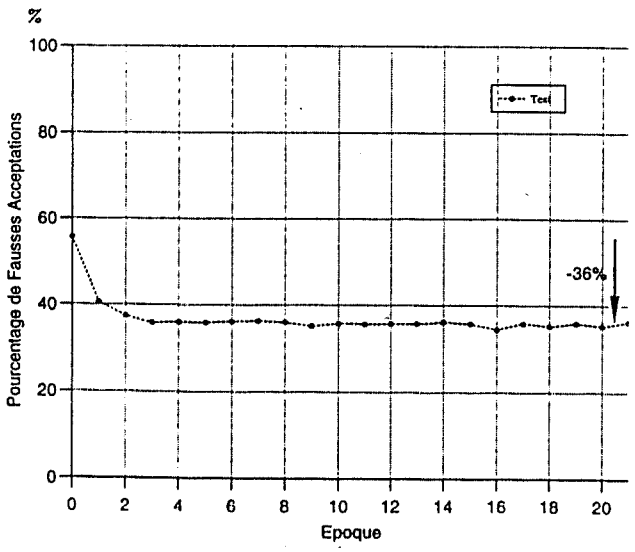


Figure 3: Rejet de la base du *Trégor*

La figure 3 montre que l'on obtient là encore une amélioration significative du taux de rejet.

5. CONCLUSION

Les résultats montrent qu'un réapprentissage discriminant afin de maximiser les critères MCE et MMI permet d'obtenir une amélioration significative des performances sur deux applications utilisant une modélisation grossière de la parole à rejeter, semblable à celle qui est utilisée dans deux serveurs vocaux en exploitation.

Un avantage supplémentaire d'une telle méthode est qu'elle n'engendre aucun surcoût durant la phase de reconnaissance.

6. BIBLIOGRAPHIE

Rose R.C., Juang B.H, et Lee C.H. (1995), A training procedure for verifying string hypothesis in continuous speech recognition, *IEEE International Conference on ASSP, Detroit, mai 1995, vol.I, pp. 281-284.*

De la Torre C. et Acero A. (1994), Discriminative training of garbage model for non-vocabulary utterance rejection, *International Conference on Spoken Language Processing, Yokohama, Japon, septembre 1994, vol.I, pp. 475-478.*

Athimon C., Bigorgne D., Cherbonnel B., Dubois D., Gagnoulet C., Jouvet D., Marzio H., Monné J., Py S., Sorin C. et. Toularhoat M. (1994), Operational and experimental french telecommunication services using CNET speech recognition and text-to-speech synthesis, *IEEE Workshop on Interactive Voice Technology for Telecommunications Applications, Kyoto, Japon, septembre 1994, pp. 27--32.*

Ney H. (1995), On the probabilistic interpretation of neural network classifiers and discriminative training criteria, *IEEE Transactions on PAMI, février 1995, vol.17, n2 pp. 107-119.*

Brown P.F. (1987), The Acoustic-Modelling Problem in Speech Recognition, *PhD thesis, Carnegie Mellon University, Pittsburgh, USA.*

.Jouvet D., Bartkova K. et Monné J. (1991) On the modelization of allophones in an HMM based speech recognition system, *2nd European Conference on Speech Communication and Technology,* Gênes, *Italie, vol. 2, pp. 923-926.*

Leprieur H. et P. Haffner (1995), Discriminant learning with minimum memory loss for improved non-vocabulary rejection, *4th European Conference on Speech Communication and Technology, Madrid, Espagne, vol.1, pp. 89-92.*

APPORTS D'UNE COMPOSANTE PHONOLOGIQUE À LA RECONNAISSANCE AUTOMATIQUE DE LA PAROLE CONTINUE

Laure Pousse - Martine de Calmès - Guy Pérennou

Institut de Recherche en Informatique de Toulouse - UPS - 118, Route de Narbonne - 61063 TOULOUSE CEDEX

Tel.: 61 55 61 73 - Fax: 61 55 62 58 - e-mail: pousse@irit.fr

ABSTRACT

This paper presents the problem of multiple pronunciations of a single word in automatic continuous speech recognition. We present the MHAT model, which solves the problems of liaisons and reductions very common in French, by using phonological and phonetic rules. Before implementing this model, we have tested the advantadges of introducing multiple pronunciations for a single orthographic word in the training and in the test lexica. We deliver here the results of this first experimentation.

1. INTRODUCTION

L'extrême variabilité des prononciations d'un même mot en fonction du locuteur ou de son contexte dans la phrase pose un important problème en reconnaissance automatique de la parole continue. On est donc en droit de penser que la phonologie joue un rôle important. Pourtant, dans la majorité des systèmes grand vocabulaire actuellement commercialisés comme les machines à dicter DragonDictate ou Speech Server d'IBM, la phonologie est quasi-inexistante.

Le but de cet article est triple : nous montrons d'abord les limites des systèmes qui ne tiennent pas compte des phénomènes phonologiques ; nous proposons ensuite un modèle permettant de pallier ces problèmes ; enfin, nous présentons une première évaluation de la prise en compte des multiples variantes de prononciation d'un même mot par une expérimentation utilisant des corpus de parole téléphonique dans une application indépendante du locuteur portant sur un vocabulaire de 1515 mots orthographiques.

2. LE PROBLÈME

Les systèmes markoviens de reconnaissance automatique de la parole continue utilisent un modèle de langage à états, où un état est responsable de l'émission d'un mot. Que ce soit dans des systèmes bigrammes ou trigrammes, un état n'émet qu'un seul mot. Comment alors prendre en compte les différentes prononciations d'un même mot, lorsque plusieurs de ces prononciations sont dépendantes du contexte ?

Prenons l'exemple du mot "grande". Il possède 3 prononciations différentes :

R1 : grãd
R2 : grãdœ
R3 : grãn

Une première approche consiste à considérer trois entrées pour le mot grande dans le lexique : grande1, grande2, et grande3. A chaque couple orthographe-prononciation est associée une probabilité indépendante du contexte. L'avantage d'un tel modèle est que l'on peut reconnaître des mots qui ne seraient pas apparus faute d'une bonne association orthographe-prononciation. Les inconvénients sont d'une part l'accroissement excessif de la taille du lexique, et d'autre part la non prise en compte du contexte du mot dans la phrase lors de la reconnaissance pour le choix de telle ou telle prononciation.

Pour pallier ce second inconvénient, nous cherchons donc à ce que le couple orthographe-prononciation ait une probabilité dépendante du contexte. Le traitement algorithmique du décodage est facilité si l'émission d'un mot ne dépend que de l'état émetteur. C'est ce que l'on s'impose dans le modèle MHAT (Markovian Harmonic Adaptation and Transduction) (Pérennou, 1995). Si par exemple l'état émetteur est un bigramme anticipateur (O1,O2) l'émission sera une des n prononciations Ri, i=1..n, du premier mot (O1) dans le contexte du second (O2) avec les probabilités p(Ri / O2), i=1..n. (Figure 1). En fait, cette prononciation ne dépend que du contexte initial de O2, et de la frontière séparant O1 et O2 (voir le paragraphe 4).

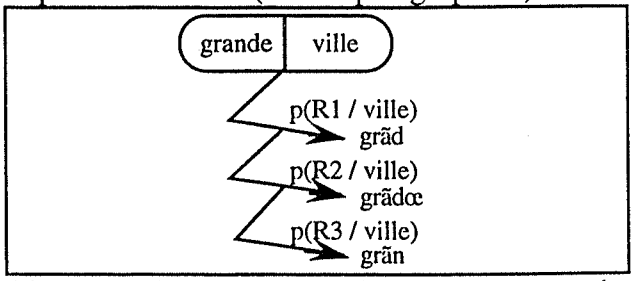


Figure 1: Un exemple de bigramme émetteur de plusieurs prononciations.

Cela revient à ce que chaque bigramme soit responsable de l'émission d'un mot complexe

sous forme phonotypique (Figure 2). Les probabilités de transformation en telle ou telle autre variante de prononciation ne dépendent plus que des lois de prononciation des groupes à prononciations multiples (gpm) qui y figurent (du gpm (~də) dans l'exemple de la figure 2). En choisissant cette représentation phonotypique, on limite la prolifération des formes de mots.

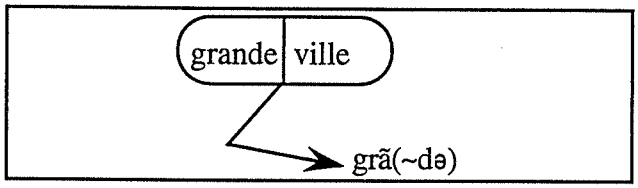


Figure 2: Un exemple de bigramme émetteur d'un mot sous forme phonotypique.

3. ÉTAT ACTUEL DES SYSTÈMES PAR RAPPORT À CE PROBLÈME

Beaucoup de systèmes considèrent une seule prononciation par mot. Ceux qui envisagent plusieurs prononciations par mot ne relient pas la prononciation au contexte sauf dans des cas simples, comme par exemple l'article "les" prononcé /lez/ devant une voyelle ou /le/ devant une consonne (cette approche est abordée au LIMSI (Gauvain, 1993)).

Le français soulève des problèmes particuliers à cet égard. Les traitements contextuels tels que liaisons ou réductions existent aussi pour d'autres langues comme l'anglais, mais elles ont alors des conséquences moins significatives sur les modèles des mots requis qu'en français. En anglais par exemple, les variations allophoniques peuvent être traitées par des lois multi-gaussiennes ou l'utilisation de bons allophones. Par exemple, la variabilité décrite par Cohen et Mercer (Cohen, 1975) du son /a/ dans le mot "telegraf" peut être résolue par allophones.

En français, le problème est plus important qu'en italien ou en anglais à cause des phénomènes d'élision de finales ou de liquides, qui peuvent être combinés avec l'assimilation de voisement ou de nasalité. En ce qui concerne les résultats qui attestent de l'apport d'une composante phonologique pour le français, on peut se référer aux travaux de Philips (Dugast, 1995) ou du LIMSI (Gauvain, 1993).

4. NOTRE APPROCHE

Elle vise à traiter le problème donné dans le paragraphe précédent. Une possibilité est d'utiliser le modèle MHAT : Markovian Harmonic Adaptation and Transduction (Pérennou, 1995). Cette approche consiste à décrire les variantes de prononciation et leur dépendance au contexte à travers trois séries de règles qui opèrent aux différents niveaux du modèle MHAT :

1°- au niveau syntaxique en posant les frontières existant entre les mots du point de vue des contraintes de prononciations. Les règles se présentent sous la forme de contraintes biclasses, comme cela est illustré ci-après (Table 1).

Table 1: Exemple de règles de pose de frontières (voir (Table 2) pour des exemples de classes lexicales)

no de règle	Partie gauche	Frontière	Partie droite	Exemples
F2	Vf, Vinf, Vant	≠ facul-tative	V,p,d,J,A	il prit ≠ un livre
F4	d, J1	# oblig-atoire	N,J,A,Pi	un # ours

Table 2: Exemples de classes lexicales

N : nom	J : adjectif
Vf : verbe conjugué (fini)	V inf : verbe à l'infinitif
Vant : verbe au participe présent	p : préposition
d : déterminant	J : adjectif
A : adverbe	J1 : adjectif antéposé à la liaison facultative
N : nom	Pi : pronom indéfini

Ce type de règles est parfois markovien d'ordre 1 (il ne met en jeu que 2 mots consécutifs). Dans bien des cas, il est de type monogramme (il ne met en jeu qu'un seul mot). Ces frontières jouent un rôle important pour le traitement des liaisons et dans une certaine mesure, pour le traitement du schwa.

2°- Les règles d'adaptation contextuelles (ou encore règles d'adaptation harmonique (Pérennou, 1995)) permettent de prendre en compte le contexte soit initial soit terminal du mot ; les règles se présentent sous la forme :

< ~də > —> d / _{#, ≠} NC
(exemple : [grɑ̃durs] grande ourse)
—> (~də) / _{#, ≠} OBM
(exemple : [grɑ̃nmoto] grande moto)
—> (də) / sinon
(exemple : [ɛlegrɑ̃dœ] elle est grande)

avec NC = non consonantique
OBM = occlusives, fricatives, liquides, nasales

Ces règles procèdent de manière irrémédiable en un seul pas. Elles sont du type bigramme (autrement dit elles ne dépendent que de deux mots consécutifs).

3°- Les règles de prononciation qui s'expriment de la manière suivante :

(~də) -> dœ | d | n
(də) -> dœ | d

Ces règles permettent de rendre compte des variations de prononciation d'un même mot dans un même contexte.

Ces règles sont probabilistes et sont ajustées par apprentissage sur des corpus (Cf. travaux d'Alix Mailland (Mailland, 1995) sur BREF (Lamel, 1991)).

Nous disposons de ces matériaux et de plus, d'un lexique LexMHAT donnant les représentations en gpc (groupes phonologiques contextuels) et en gpm (groupes à prononciations multiples) des mots.

Des outils permettent de générer les prononciations des mots à plusieurs niveaux :

- au niveau W' (représentation phonotypique (Table 3))
- au niveau P (représentation en allophones (Table 4))

Table 3: Représentations phonotypiques (en gpm)

Orthographe	Représentation W'	Contexte
grande	grã(d)	_{#,≠} NC
grande	grã(~də)	_{#,≠} OBM
grande	grã(də)	sinon

Table 4: Représentation en allophones des prononciations

Orthographe	Représentation P	Contexte
grande	R1	tous contextes
grande	R2	non(_ {#,≠}NC)
grande	R3	_{#,≠} OBM

5. UNE EXPÉRIENCE

L'expérimentation du modèle MHAT n'est intéressante que sur de grands corpus. Nous avons commencé par faire des statistiques sur les gpm et gpc présents dans le vaste corpus BREF (on peut se référer à (Mailland, 1995) pour les résultats).

La validation de ce modèle MHAT nécessite un système assez puissant pour traiter la parole continue indépendante du locuteur avec des vocabulaires assez importants.

Dans un premier temps, nous avons comparé les taux de reconnaissance avec plusieurs variantes de prononciations et sans variante (en utilisant d'abord la prononciation la plus longue, puis la plus courte), sans faire intervenir le contexte dans lequel les variantes étaient interdites ou autorisées, ceci dans le but de vérifier l'apport d'une composante phonologique dans un système de reconnaissance automatique de la parole continue. Puis, nous avons effectué la même comparaison lorsque l'apprentissage est modifié de la même manière (avec et sans variante de prononciation). Enfin, la dernière comparaison a été effectuée en utilisant en apprentissage une seule variante de prononciation, alors que la reconnaissance garde toutes les variantes.

Le corpus sur lequel ont été effectuées ces expériences est un corpus d'une vingtaine de locuteurs. Il contient 15795 mots en 2322 phrases reparties entre un corpus d'apprentissage (13504 mots, 1910 phrases) et un corpus de test (2291 mots, 412 phrases). Les lexiques utilisés comportent 1515 graphies, ce qui correspond à 3717 entrées lorsqu'on utilise plusieurs prononciations pour un même mot (on a en moyenne 2,45 prononciations différentes par forme fléchie).

Les locuteurs et les phrases utilisés pour l'apprentissage ne sont pas dans le corpus de test. Pour le corpus de test, les locuteurs ont été laissés libres de leurs interventions, à condition qu'ils restent dans le domaine de l'application, qui était l'interrogation d'une base de données relatives à des renseignements ferroviaires. De plus, nous souhaitions nous mettre dans les conditions les plus réalistes de telles interrogations. Nous avons donc enregistré ce corpus dans le bruit ambiant d'une salle machine, et nous avons rééchantillonné le signal pour obtenir la "qualité" téléphonique.

Le modèle de langage bigramme utilisé a été obtenu à partir de différentes phrases propres au domaine (dont les phrases du corpus d'apprentissage). Les densités utilisées sont des mixtures de laplaciens. Compte tenu de la faible puissance des machines utilisées et de la petite taille des corpus, nous avons travaillé avec des phonèmes hors contexte. A l'heure actuelle, l'utilisation de triphones pourrait faire descendre les taux d'erreurs autour de 15%. Cependant, ce que nous recherchons c'est la mesure de la contribution de la phonologie et cette expérience demeure significative à cet égard.

6. RÉSULTATS

Expérience pilote : En apprentissage, le lexique contient la variante de prononciation effectivement utilisée (la bonne). En reconnaissance, le lexique contient toutes les variantes de prononciations possibles pour chaque mot orthographique.

Première expérience : En apprentissage le lexique contient la bonne prononciation. En reconnaissance le lexique ne contient qu'une seule variante de prononciation par mot orthographique.

Table 5: Comparaison des taux d'erreurs sur les mots entre l'expérience pilote et la première expérience.

Expérience	Variante de prononciation		Taux d'erreurs (%) décembre 95
	A	R	
pilote	la bonne	toutes	25,80 % 377s 60i 154d
Expérience 1	la bonne	la plus longue	36,88 % 499s 39i 307d
Expérience 1	la bonne	la plus courte	33,35 % 500s 78i 186d

Deuxième expérience : En apprentissage, comme en reconnaissance, le lexique ne contient qu'une seule variante de prononciation par mot orthographique.

Table 6: Comparaison des taux d'erreurs sur les mots entre l'expérience pilote et la deuxième expérience.

Expérience	Variante de prononciation		Taux d'erreurs (%) décembre 95
	A	R	
pilote	la bonne	toutes	25,80 % 377s 60i 154d
Expérience 2	la plus longue	la plus longue	36,84 % 515s 46i 283d
Expérience 2	la plus courte	la plus courte	30,73 % 444s 71i 189d

Troisième expérience : En apprentissage le lexique contient une seule variante de prononciation par mot orthographique, alors qu'en reconnaissance, le lexique contient toutes les variantes de prononciations possibles.

Table 7: Comparaison des taux d'erreurs sur les mots entre l'expérience pilote et la troisième expérience.

Expérience	Variante de prononciation		Taux d'erreurs (%) décembre 95
	A	R	
pilote	la bonne	toutes	25,80 % 377s 60i 154d
Expérience 3	la plus longue	toutes	32,13 % 480s 73i 183d
Expérience 3	la plus courte	toutes	27,54 % 419s 55i 157d

• Le taux d'erreurs sur les mots passe de 25,80% à 33,35%, lorsqu'on ne tient plus compte de plusieurs variantes de prononciations. On améliore donc la reconnaissance en augmentant le nombre de variantes de prononciations, malgré l'augmentation de la taille du lexique.

• Le taux d'erreurs de reconnaissance des mots est de 36,88 % en utilisant la variante de prononciation la plus longue , alors qu'il n'est que de 33,35% lorsqu'on utilise la variante la plus courte. Cette différence met en avant la grande importance du bon choix de la variante de prononciation utilisée, lorsqu'il n'y en a qu'une. Pour éviter de prendre une mauvaise décision dans ce choix, le mieux est donc de proposer plusieurs variantes de prononciations d'un même mot.

• Comme on s'y attendait, les résultats sont meilleurs (30,73% contre 33,35% et 36,84% contre 36,88%) lorsque l'apprentissage est effectué sur le même modèle que la reconnaissance.

• La reconnaissance des mots est toujours meilleure lorsqu'on utilise un lexique avec des variantes de prononciations (même si l'apprentissage n'a utilisé qu'une seule prononciation par mot). Les scores sont de 32,13% contre 36,84% et de 27,54% contre 30,73%.

7. CONCLUSION ET PERSPECTIVES

Les tests que nous avons réalisés sont évidemment en deçà du modèle MHAT car ils ne tiennent pas compte du contexte des mots : les émissions des variantes de prononciations d'un même mot sont équiprobables. Néanmoins, ils permettent déjà de mettre en évidence le rôle de la phonologie dans un système de reconnaissance de la parole continue.

Les expériences actuelles visent à prendre en compte l'effet du contexte conformément à MHAT, ce qui nous conduit à redéfinir les modèles de langage au niveau phonotypique, que ce soit pour la préparation des corpus d'apprentissage des modèles de langage ou pour la reconnaissance.

8. BIBLIOGRAPHIE

Pérennou G. (1995) MHAT : une composante phonologique pour la reconnaissance automatique de la parole. Le cas des consonnes latentes en français, *Rapport interne IRIT 95-23-R*

Gauvain J.-L., Lamel L.F., Adda G., Adda-Decker M. (1993) Speaker-Independent Continuous Speech Dictation, *Eurospeech93*, 125-128

Cohen P.S. & Mercer R.L. (1975) The Phonological Component of an Automatic Speech-Recognition System, *Speech Recognition*, Academic Press New-York, 275-320

Dugast C., Aubert X., Kneser R. (1995) The Philips large-vocabulary recognition system for American English, French and German, *Eurospeech95*, 197-200

Mailland A., de Calmès M., Pérennou G. (1995) Learning of a Phonological Component from Bref Corpus, *ICPhS Stockholm*, 610-613

Lamel L.F., Gauvain J.-F., Eskénazi M. (1991) BREF, a Large Vocabulary Spoken Corpus for French, *Eurospeech91*, 505-508

Approche neuromimétique en traitement acoustique. Le modèle TOM

Stéphane DURAND, Frédéric ALEXANDRE

Crin-CNRS / Inria Lorraine - BP 239, 54500 Vandœuvre-lès-Nancy

Tel.: 83 59 20 53 - Fax : 83 41 30 79 - e-mail: durand@loria.fr, falex@loria.fr

ABSTRACT

In this paper, we present a temporal connectionist system based on neuro-biological data. The temporal dimension, involved in speech recognition problems, is dealt within the network. The network is able to learn and recognize acoustic information sequences and is tolerant to temporal distorsion. The results obtained on spoken digits recognition are close to the ones obtained with Markov models.

1 INTRODUCTION

Notre objectif consiste en la conception d'un modèle connexionniste temporel dont le principe est fondé sur la propagation d'activité au sein d'un ensemble d'entités connexionnistes. Cette propagation conduit à l'apprentissage et la reconnaissance de séquences de stimuli avec une tolérance aux distorsions temporelles. Les entités neuronales sont organisées sous la forme de cartes que nous appellerons TOM (Temporal Organization Map). Dans les applications, une carte est dédiée à un traitement spatio-temporel donné avec une échelle de temps donnée. En particulier, en reconnaissance de la parole une carte de bas niveau peut apprendre et stocker de petites séquences acoustiques correspondant à des phonèmes ou des syllabes. Une telle carte joue deux rôles : elle réagit à des stimuli acoustiques et ses éléments connexionnistes réagissent conditionnellement à des contextes ou des passés d'activation particuliers.
Dans sa formalisation, le modèle de l'entité connexionniste qui compose ce genre de carte se rapproche plus du modèle de la colonne corticale (Burnod, 1988) que du modèle du neurone formel de MacCulloch et Pitts. Dans sa philosophie et aussi du point de vue des résultats, une comparaison peut être amenée vers les modèles stochastiques classiquement utilisés en reconnaissance de la parole : les HMM.
Notre modèle est structurellement et fonctionnellement générique et ne dépend pas de l'application. Il peut exploiter les propriétés des réseaux de neurones statiques par la fusion multi-capteurs notamment. Cependant, dans cet article, nous nous limitons à une application de reconnaissance de chiffres parlés en mode multi-locuteurs.

2 THEORIE ET FONCTIONNEMENT DU MODELE TOM

2.1 Architecture

Une carte TOM est composée d'un ensemble de *super-unités* que l'on définit ainsi parce qu'elles correspondent à des unités fonctionnelles plus complexes que le neurone utilisé dans les perceptrons ou les cartes auto-organisatrices. Différents types de liens aboutissent sur ces super-unités, chacun d'eux possède une fonctionnalité et une utilité particulière (fig. 1).

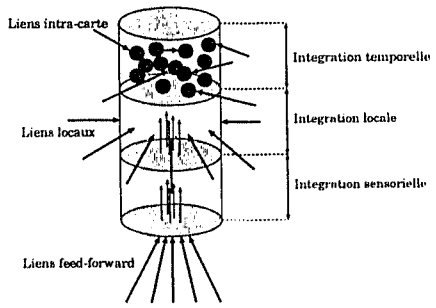


FIG. 1 – L'élément fondamental d'une structure TOM : la super-unité

La fonctionnalité spatiale est assurée par des liens feed-forward et par des liens locaux ou

de voisinage. La fonctionnalité temporelle est assurée par les *unités* inclusent dans les super-unités et par les liens intra-carte qui assurent la connectivité entre ces unités.

2.2 Principes de fonctionnement

2.2.1 *Codage des stimuli, influence du voisinage*

Les stimuli sont représentés par des vecteurs et sont codés au niveau d'une carte TOM par les vecteurs de poids feed-forward des super-unités. Ainsi, après apprentissage, la présence d'un stimulus déclenche l'activation (binaire) d'une super-unité.
Les liens de voisinage (ou locaux) définissent une topologie à la manière des cartes de Kohonen. Ceci permet en outre une tolérance aux variabilités spatiales et la reconnaissance de séquences proches de séquences codées sur la carte (figure 2).

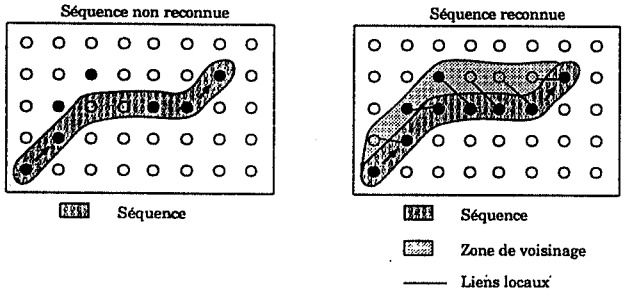


FIG. 2 – Séquence et voisinage. A gauche, pas de voisinage, à droite, avec voisinage

2.2.2 *Activation des unités*

L'activation d'une unité est déclenchée par la présence simultanée du stimulus associé à la super-unité dans laquelle se trouve l'unité ($SA = 1$) et d'un stimulus dans le champ récepteur temporel (composé de liens intra-carte) de l'unité ($TA = 1$). L'expression suivante donne la valeur de déclenchement :

$$\overline{Act(t)} = \frac{1}{\alpha + \beta}(\alpha SA + \beta TA) \qquad (1)$$

où les coefficients α et β règlent l'influence respective de chacune des composantes, spatiale ou temporelle.
Pour propager une activité, il est nécessaire de la « mémoriser » durant un laps de temps. C'est la raison pour laquelle les activités des unités décroissent au cours du temps selon un paramètre de décroissance *decay* qui contrôle le maintient d'une forte activité plus ou moins longuement. L'activité de l'unité est alors définit de la façon suivante :

$$Act(t) = \begin{cases} 1 & \text{si } \overline{Act(t)} \geq 1 \\ decay.Act(t-1) & \text{sinon} \end{cases} \qquad (2)$$

La figure 3 montre un exemple d'activation d'une unité au cours du temps.

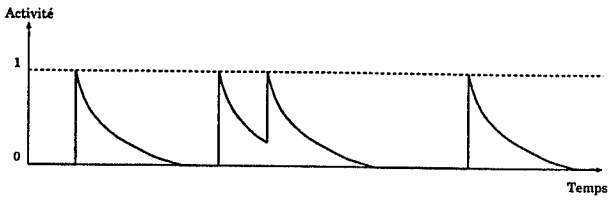


FIG. 3 – Loi d'activation d'une unité

2.2.3 *Propagation et reconnaissance*

Une forme dynamique (une séquence d'information acoustique par exemple) possède un modèle de propagation d'activité qui lui est propre. En phase de reconnaissance, il s'agit de voir si la propagation qui se produit correspond à un modèle de propagation codé sur la carte. En d'autres termes, il s'agit de regarder si des unités fin de séquence s'activent.

2.3 Apprentissage

L'apprentissage s'effectue séquentiellement en deux étapes. Il y a d'abord construction des champs récepteurs spatiaux (liens feed-forward), puis construction des liens intra-carte pour l'apprentissage des combinaisons temporelles (séquences).
Pour la première phase (spatiale) nous utilisons classiquement un algorithme basé sur les cartes auto-organisées de type carte de Kohonen (Kohonen, 1984).
Par manque de place, nous ne pouvons donner et détailler ici l'algorithme d'apprentissage des liens intra-carte (consulter (Durand, 1995) pour cela). La figure 4 résume schématiquement la construction de tels liens pour une séquence de lettres *BCDAB*(à une super-unité correspond une lettre).

Un paramètre de profondeur de connexion contrôle la création des liens vers des unités plus ou moins lointaines dans le temps.
Parce que l'apparition d'un stimulus peut se faire dans diverses conditions, en particulier

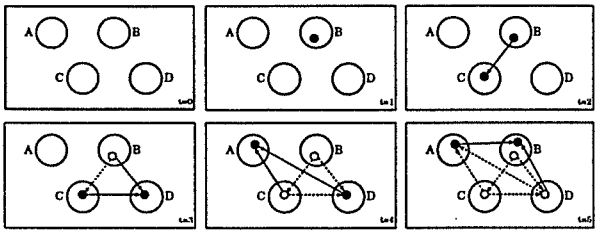


FIG. 4 – Apprentissage d'une séquence, évolution sur la carte

dans divers contextes temporels, il nous faut pouvoir distinguer les stimuli à l'information spatiale identique mais au passé différent. C'est la raison pour laquelle, à chaque fois qu'un nouveau contexte d'activation est connu pour un stimulus donné, nous créons une nouvelle unité à l'intérieur de la zone d'intégration temporelle de la super-unité.

2.4 Tolérance temporelle sélective

Dans ce qui précède, chaque événement a la même importance dans la séquence (les paramètres de mémoire à court terme et et de profondeur de connexion sont valables pour tout le réseau et sont identiques pour chaque unité). Or, dans la réalité, un événement n'a pas toujours la même importance suivant sa position dans la séquence ou suivant sa fréquence d'apparition dans le corpus.
Par exemple en parole, si l'on considère une syllabe composée d'une occlusive suivie d'une voyelle, l'information caractérisant l'occlusive se trouve confinée dans un intervalle de temps réduit par rapport à l'intervalle de temps utile pour la voyelle. Pour cette dernière, une tolérance temporelle assez large peut intervenir alors que pour l'occlusive il n'est pas possible de tolérer des suppressions et des insertions d'événements.
Afin de tenir compte de ce fait, nous assignons des paramètres de décroissance d'activation et de profondeur de connexion propres à chaque unité. Ces paramètres seront alors appris en fonction de trois cas de configuration.
Dans le premier cas, une séquence est présentée au réseau. Il s'en suit une propagation d'activité relativement faible et n'aboutissant pas au déclenchement d'une séquence codée. Nous considérons donc la séquence présentée comme nouvelle et la codons directement dans la carte en lui affectant des valeurs de tolérance (*decay* et *thres_act*) définies *a priori*.
Dans le second, une séquence est présentée au réseau et induit une propagation d'activité conduisant à une reconnaissance correcte. La séquence ayant conduit à cette reconnaissance est repérée et fusionnée avec la séquence d'entrée. Les unités communes à ces deux séquences sont considérées comme étant les unités principales de la séquence (responsables de son activité), c'est la raison pour laquelle les paramètres de tolérance sont adaptés pour forcer leur présence dans les activations futures. Une façon de renforcer l'importance de ces unités est d'augmenter la décroissance pour chacune d'elle et de diminuer la profondeur de connexion pour les autres.
Enfin dans le dernier, une séquence est présentée au réseau et induit une propagation d'activité aboutissant à une confusion dans la reconnaissance Dans ce cas, il est nécessaire de « séparer » les deux séquences. Si les deux séquences ont été activées c'est que des unités communes l'ont été. Il convient d'empêcher par la suite l'activation de ces unités communes. La technique employée ici consiste à diminuer les deux paramètres décroissance et profondeur.

3 EXPERIMENTATION

Nous présentons ici les résultats obtenus sur des corpus de chiffres parlés avec une carte à paramètres de distorsions locaux et appris.
Nous disposons d'une carte d'apprentissage des séquences acoustiques connectée à une carte de détection de séquence qui étiquette en quelque sorte les séquences codées. Les résultats sont obtenus à partir de corpus issus de la base de données de chiffres américains TIDIGIT. Deux expériences ont été menées. La première, qualifiée de « pseudo multi-locutrice » comporte un corpus d'apprentissage d'une dizaine de locuteurs prononçant chacun 10 chiffres (100 chiffres au total) et un corpus de test contenant les 10 mêmes locuteurs prononçant d'autres occurrences des chiffres (100 chiffres). La seconde est complètement multi-locutrice. Elle comporte un corpus d'apprentissage de 56 locuteurs prononçant chacun deux occurrences de chaque chiffre (1120 chiffres) et un corpus de test de 17 autres locuteurs (Durand, 1995 ; Durand 1996).

Type test	Train	Test
Pseudo multi-locuteur	100,00%	100,00%
Multi-locuteur	98,75%	98,52%

Ces résultats ont été obtenus avec des liens feed-forward et de voisinage définissant une carte de Kohonen à deux dimensions comportant 400 super-unités. Les paramètres initiaux de profondeur de connexion et de décroissance d'activation valent respectivement 0.63 et 0.96. Chaque super-unité sur la carte de détection de séquences est connectée à la totalité des unités qui compose la séquence. elle réagit au pourcentage d'unités active le long de la séquence.

3.1 Comparaison avec les modèles de Markov

Ces résultats, même s'ils demandent a être plus poussés, sont intéressants d'une part dans la mesure où l'architecture qui est mise en œuvre se détache des architectures classiques d'autre part parce qu'ils se rapprochent des résultats obtenus par les meilleurs systèmes dans ce domaine comme les HMM (Hidden Markov Model). Ceux-ci semblent bien modéliser le phénomène temporel. Ils sont d'ailleurs souvent couplés à des réseaux de neurones pour exploiter conjointement les qualités des deux techniques, le modèle de Markov se chargeant de l'aspect temporel.
Le tableau 1 donne les résultats obtenus pour des modèles de Markov d'ordre 1 et 2 avec 9 ou 11 boucles et 55 locuteurs d'apprentissage comme dans le test précédent. Les chiffres sont donnés pour 56 locuteurs de tests.

TAB. 1 – Résultats obtenus avec des modèles de Markov

	9 boucles	11 boucles
HMM ordre 1	98,62%	98,70%
HMM ordre 2	98.70%	99,26%

En termes de performance, notre modèle peut donc être comparé aux HMM. Du point de vue du modèle, les deux approches tentent toutes deux d'apprendre des séquences d'information avec une tolérance aux distorsions temporelles. Les méthodes employées diffèrent cependant : elles sont statistiques pour les HMM et d'inspiration neurobiologique pour le modèle TOM.

4 CONCLUSION

Le modèle TOM, présenté ici, possède selon nous deux intérêts généraux. Tout d'abord, nous avons vu les résultats intéressants obtenus sur des tests de reconnaissance de chiffres parlés. Certes ces résultats ne dépassent pas encore les meilleurs résultats obtenus dans le domaine (notamment avec les chaînes de Markov). Il faut cependant noter que les modèles de Markov sont plus anciens et ont été extensivement étudiés et testés ce qui n'est pas le cas du modèle TOM. Ensuite, l'approche que nous utilisons est originale. A l'image du cortex cérébral, nous développons une entité connexionniste avec une fonctionnalité étendue et théoriquement indépendante de l'application. Le dernier point nous semble crucial pour les applications et les améliorations à venir. En effet, du fait de sa généricité une carte TOM peut servir à différents traitement et peut s'insérer dans divers axes sensoriels. En particulier, on pourrait imaginer de coupler une carte d'apprentissage de séquences acoustiques avec une carte chargée de détecter les mouvements de la bouche lors de la prononciation.

5 BIBLIOGRAPHIE

Burnod Y. (1988) – *An adaptive neural network: The cerebral cortex.* – Masson Paris.

Durand S. (1995) – *TOM, une architecture connexionniste de traitement de séquences. Application à la reconnaissance de la parole.* – Thèse de doctorat, Université Henri Poincaré, Nancy I.

Durand S. et Alexandre F. (1996) – TOM, a new temporal neural net architecture for speech signal processing. *In : ICASSP, IEEE International Conference on Acoustic Speech and Signal Processing.* – Atlanta, USA.

Kohonen T., Makisara K. et Saramaki T. (1984) – Phonotopic maps – insightful representation of phonological features for speech recognition. *In : IEEE Proceedings of the 7th International Conference on Pattern Recognition.*

COMBINAISON DE DIFFERENTES MODELISATIONS CONTEXTUELLES POUR LA RECONNAISSANCE FLEXIBLE

J. Simonin, S. Bodin, D. Jouvet & K. Bartkova

CNET - LAA/TSS/RCP - 2 Av. Pierre Marzin - 22307 LANNION

Tél: 96 05 23 10 - Fax: 96 05 35 30 - e-mail: simonin@lannion.cnet.fr

ABSTRACT

This article presents different contextual phonetic modelings for a speaker independent flexible (i.e. task independent) speech recognition system. Evaluations are conducted on telephone speech data in order to determine the best trade-off between computational cost and recognition performances. The allophone models lead to good recognition performances with a reduced number of parameters. However, recognition performances may be improved at the price of an increase of the computational cost, by combining allophone and triphone models.

1. INTRODUCTION

Les techniques de reconnaissance de parole doivent rendre compte de la variabilité du signal de parole et en particulier des effets contextuels, c'est-à-dire, des variations acoustiques du signal de parole dues au contexte phonétique dans lequel le signal est émis.

Ces influences contextuelles sont prises en compte par l'emploi d'unités contextuelles qui représentent les réalisations acoustiques des sons dans des contextes gauche et droit spécifiés (Schwartz, 1985). L'importance du nombre de ces unités contextuelles entraine un apprentissage impossible de la totalité des unités et une estimation pas toujours fiable des paramètres de ces unités. La réduction du nombre d'unités peut reposer sur un regroupement des modèles contextuels appris (Lee, 1990) ou sur l'utilisation d'arbres de décision afin de déterminer le modèle à choisir suivant le contexte spécifié (Bahl, 1989).

Dans le cadre d'un système de reconnaissance de la parole flexible, où aucun apprentissage spécifique à la tâche n'est effectué, une bonne modélisation des effets contextuels est nécessaire. Les unités contextuelles, dédiées à des contextes spécifiques, sont alors utilisées de telle manière que les modèles de chaque mot du vocabulaire soient dérivés d'une concaténation de telles unités.

L'objectif de cette étude est d'optimiser la modélisation en choisissant la meilleure combinaison d'unités contextuelles parmi les triphones, les diphones et les allophones. Ces unités différent au niveau du partage des paramètres et des réalisations contextuelles qu'elles modélisent et donc aussi au niveau de la fiabilité des estimations résultantes.

2. TRIPHONES ET ALLOPHONES

La prise en compte des influences contextuelles permet une meilleure modélisation acoustique des phonèmes (Schwartz, 1985). Les diverses modélisations dépendent des contextes considérés et du partage des paramètres entre les unités.

2.1. Les modèles contextuels utilisés

Un triphone est une unité contextuelle modélisant un phonème dans des contextes gauche et droit spécifiques. Chaque triphone a son propre jeu de paramètres.

Le partage des paramètres, en particulier ceux qui modélisent la partie centrale, appelée cible, entraine une réduction du nombre de paramètres, et donc une estimation plus fiable de ces paramètres lors de l'apprentissage. La modélisation allophonique (Bartkova, 1991, et Jouvet, 1994) est déduite de ce principe. Un allophone intègre au sein d'une même unité toutes les réalisations contextuelles d'un phonème. Il est constitué d'une ou plusieurs cibles, et de divers points d'entrée et de sortie correspondant aux différents contextes possibles. Il regroupe dans de mêmes points d'entrée ou de sortie des contextes,

acoustiquement proches, qui créent le même type de déformation acoustique sur le phonème considéré. Contrairement au triphone, il met donc en oeuvre un partage a priori des paramètres.

Le modèle à base de triphones est le modèle potentiellement le plus précis car le plus spécifique à un contexte donné. Mais, du fait même de sa spécificité, le triphone peut apparaître assez rarement dans le corpus d'apprentissage d'où une estimation du modèle non fiable. En revanche, la modélisation allophonique de par ses regroupements de contextes et son partage des paramètres conduit à une estimation plus fiable. La comparaison entre triphones et allophones repose finalement sur un compromis entre finesse de modélisation et fiabilité d'estimation des modèles.

2.2. Les corpus utilisés

Le corpus utilisé lors de l'apprentissage est constitué d'enregistrements téléphoniques de 683 phrases courtes par plusieurs centaines de locuteur. Ce corpus a la propriété de contenir quasiment tous les diphones.

Les différents corpus téléphoniques employés pour les évaluations sont décrits dans la table 1. Le nombre de locuteurs associé à chaque corpus, ainsi que le nombre d'enregistrements des ensembles de tests y sont répertoriés.

Table 1: Caractéristiques des corpus employés.

Nom	Nb. Loc.	Nb. Enr.
Chiffres (0 à 9)	335	3622
Trégor (36 mots)	380	12844
Nombres (00 à 99)	385	7288

3. COMBINAISON TRIPHONES ET DIPHONES

3.1. Modélisation acoustique

En mode indépendant de la tâche, tous les triphones requis pour modéliser le vocabulaire de l'application n'ont pas nécessairement été rencontrés dans le corpus d'apprentissage, ou du moins, pas suffisamment pour être bien estimés. Il devient difficile, voire impossible, de construire un modèle exclusivement à base de triphones. Dans cette optique, l'apprentissage de tous les triphones rencontrés dans le corpus d'apprentissage est effectué, mais aussi de tous les diphones droits, tous les diphones gauches ainsi que des phonèmes hors contexte. Un diphone droit (resp. gauche) modélise un phonème dans un contexte droit (resp. gauche) spécifique et un contexte gauche (resp. droit) quelconque.

Le modèle d'un vocabulaire donné est créé en modélisant les phonèmes décrivant les mots de la façon suivante. Le premier modèle correctement estimé lors de l'apprentissage parmi le modèle triphone, le modèle diphone droit et le modèle diphone gauche, dans cet ordre, est utilisé. Si aucun de ces modèles n'a été suffisamment estimé, le modèle du phonème hors contexte est employé.

Le critère de qualité d'apprentissage des modèle des unités est le nombre d'occurrences de ces unités lors de l'apprentissage. Les tests sont effectués avec différents seuils S sur le nombre minimal d'occurrences des triphones lors de l'apprentissage. Il s'agit donc de comparer un modèle uniquement à base d'allophones avec un modèle à base de triphones, diphones et phonèmes hors contexte.

3.2. Résultats

Les taux d'erreur sur chaque corpus sont présentés dans les tables 2, 3 et 4. La première colonne indique les résultats de référence du modèle allophonique et les colonnes suivantes ceux de la modélisation à base de triphones en fonction du seuil S.

Table 2: Taux d'erreur sur le corpus des Chiffres.

Modèle	*All*	S=200	S=50	S=25	S=1
Nb.Gauss	*263*	402	430	444	452
Err. (%)	*4,22*	4,56	3,57	3,74	3,70

Table 3: Taux d'erreur sur le corpus du Trégor.

Modèle	*All*	S=200	S=50	S=25	S=1
Nb.Gauss	*842*	1590	1904	2112	2191
Err. (%)	*2,10*	3,85	2,82	3,69	3,52

Table 4: Taux d'erreur sur le corpus des Nombres.

Modèle	*All*	S=200	S=50	S=25	S=1
Nb.Gauss	*514*	876	1136	1287	1465
Err. (%)	*8,20*	10,70	9,54	9,41	10,10

3.3. Commentaires

Lorsque le seuil est trop faible, le modèle utilise beaucoup de triphones, potentiellement mal appris, d'où une dégradation sensible des performances. Inversement, si le seuil est trop grand, le modèle n'utilise que des unités bien apprises, mais moins spécifiques du contexte, même à la limite des phonèmes non contextuels. C'est pourquoi il apparaît entre ces deux extrêmes un seuil optimal correspondant à un jeu d'unités les plus détaillées possibles tout en étant correctement apprises. Sur les corpus testés, il est de l'ordre de 25 à 50.

Dans le meilleur des cas, le modèle à base de triphones, diphones et phonèmes hors contexte aboutit à des dégradations de performances par rapport aux allophones de 34% et de 15%, respectivement sur les corpus du Trégor et sur celui des Nombres. Sur le corpus des Chiffres, il conduit à une amélioration de 15% par rapport aux allophones.

Le modèle à base de triphones est donc plutôt moins efficace que le modèle à base d'allophones, tant sur le plan des performances que sur le plan du nombre de paramètres utilisés. En effet, le modèle d'allophones, grâce à ses stratégies de regroupements de contextes, est beaucoup plus léger (2 à 3 fois moins de paramètres). Seul le corpus des Chiffres est sujet à une amélioration car il est en bonne adéquation avec le corpus d'apprentissage. Il s'avère en effet que la plupart des mots du corpus des Chiffres sont présents dans le corpus d'apprentissage, ce qui assure une bonne estimation de leurs triphones.

La table 5 souligne cette particularité du corpus des Chiffres en indiquant pour chaque corpus le nombre cumulé de triphones en fonction de leur nombre d'occurences dans le corpus d'apprentissage. Le cumul signifie que chaque unité est comptée autant de fois qu'elle apparait dans la description du corpus considéré. Le nombre cumulé de triphones ayant un nombre d'occurences inférieur à 20, 30 et 50 dans le corpus d'apprentissage est comparé au nombre cumulé de triphones du corpus étudié.

Table 5: Nombres cumulés de triphones en fonction du nombre d'occurences observées lors de l'apprentissage.

Corpus	*Cumul triph.*	Nombre d'occurences		
		<20	<30	<50
Chiffres	*57*	1	2	9
Trégor	*724*	66	80	218
Nombres	*7942*	929	1419	3145

3.4. Choix du ``meilleur'' diphone

Dans le modèle précédent, quand le triphone est mal estimé, c'est-à-dire, lorsqu'il y a lieu d'utiliser le diphone, le diphone droit est sélectionné en priorité par rapport au diphone gauche, et ce choix arbitraire est d'autant moins judicieux que le diphone gauche est mieux estimé. L'idée développée ici consiste à reprendre les tests précédents, en remplaçant tous les diphones droits par les diphones gauches quand ceux-ci sont mieux estimés que les diphones droits d'après le critère du nombre d'occurrences d'apprentissage .

Les tables suivantes, 6 et 7, comparent les résultats obtenus sur les corpus des Chiffres et du Trégor, pour le modèle à base d'allophones et le modèle résultant de la sélection du diphone, droit ou gauche, le mieux appris.

Table 6: Taux d'erreur sur le corpus des Chiffres.

Modèle	*All*	**S=200**	**S=50**	**S=25**	**S=1**
Nb.Gauss	*263*	409	437	444	452
Err. (%)	*4,22*	4,39	3,53	3,65	3,65

Table 7: Taux d'erreur sur le corpus du Trégor.

Modèle	*All*	**S=200**	**S=50**	**S=25**	**S=1**
Nb.Gauss	*842*	1568	1954	2134	2191
Err. (%)	*2,10*	4,37	3,20	4,09	4,25

Par rapport à la modélisation précédente, l'amélioration est infime sur le corpus des Chiffres, et la dégradation est assez nette sur le corpus du Trégor. Les diphones droits et gauches ayant été estimés séparément, il est vraisemblable qu'ils ont subi des dérives différentes lors de l'apprentissage, et qu'un modèle combinant les deux types de diphones devient très sensible à ces défauts d'alignement.

4. COMBINAISON TRIPHONES ET ALLOPHONES

Les modèles à base de triphones, diphones et phonèmes hors contexte n'ayant pas apporté d'améliorations significatives par rapport aux modèles à base d'allophones, une combinaison des modélisations à base de triphones et d'allophones est étudiée.

4.1. Principe

Le modèle d'un vocabulaire donné est alors créé en modélisant les phonèmes de la manière suivante. Le modèle du triphone est utilisé s'il a été correctement estimé pendant la phase d'apprentissage. Sinon, le modèle de l'allophone correspondant est utilisé.

Chaque mot du vocabulaire est maintenant décrit à partir d'une succession alternée de triphones et d'allophones. Cette modélisation mixte tend à mettre en oeuvre la meilleure unité possible suivant la configuration d'apprentissage rencontrée. Le critère de bonne estimation reste le nombre d'occurrences d'apprentissage.

4.2. Résultats

Les tables 8, 9 et 10 présentent les résultats des évaluations sur chaque corpus.

Table 8: Taux d'erreur sur le corpus des Chiffres.

Modèle	*All*	S=200	S=50	S=25	S=1
Nb.Gauss	*263*	293	347	357	357
Err. (%)	*4,22*	3,23	2,68	2,94	3,15

Table 9: Taux d'erreur sur le corpus du Trégor.

Modèle	*All*	S=200	S=50	S=25	S=1
Nb.Gauss	*842*	1052	1501	1721	1849
Err. (%)	*2,10*	2,11	1,85	2,61	2,79

Table 10: Taux d'erreur sur le corpus des Nombres.

Modèle	*All*	S=200	S=50	S=25	S=1
Nb.Gauss	*514*	626	844	1036	1264
Err. (%)	*8,20*	7,88	7,98	9,32	9,97

Les résultats sont très positifs sur le corpus des Chiffres (jusqu'à 37% de réduction du taux d'erreur). L'amélioration est moins importante sur les corpus du Trégor (12% d'amélioration) et des Nombres (6%). La nature des résultats sur tous ces corpus semble principalement liée à la couverture en triphones du corpus d'apprentissage.

4.3. Augmentation du corpus d'apprentissage

Le corpus d'apprentissage initial a été enrichi par l'addition d'enregistrements complémentaires permettant une meilleure couverture des triphones. Les résultats obtenus avec les modèles correspondants sont présentés dans les tables 11, 12 et 13.

Table 11: Taux d'erreur sur le corpus des Chiffres.

Modèle	*All*	S=300	S=100	S=50	S=25
Nb.Gauss	*263*	346	347	354	355
Err. (%)	*3,21*	2,24	2,15	2,18	2,24

Table 12: Taux d'erreur sur le corpus du Trégor.

Modèle	*All*	S=300	S=100	S=50	S=25
Nb.Gauss	*892*	1816	1917	1946	1966
Err. (%)	*1,53*	1,44	1,41	1,40	1,47

Table 13: Taux d'erreur sur le corpus des Nombres.

Modèle	*All*	S=300	S=100	S=50	S=25
Nb.Gauss	*528*	959	1078	1157	1269
Err. (%)	*7,50*	6,79	6,36	6,29	6,99

Les résultats sur les différents corpus sont favorables quelque soit la valeur du seuil employé. L'amélioration optimale est de 33% sur le corpus des Chiffres, de 9% sur le corpus du Trégor et de 16% sur celui des Nombres par rapport aux taux d'erreurs obtenus avec une modélisation par allophones.

Les derniers résultats montrent clairement l'importance de la couverture des triphones par le corpus d'apprentissage et l'intérêt du modèle allophonique en temps que complément au modèle de triphones.

5. CONCLUSION

La combinaison d'un modèle basé sur des triphones et d'un modèle basé sur les diphones est moins efficace qu'une combinaison d'un modèle basé sur des triphones et d'un modèle allophonique. De plus, l'augmentation de la couverture des triphones par l'élargissement du corpus d'apprentissage conduit à une amélioration intéressante des performances de reconnaissance en mode indépendant de la tâche.

Les expériences montrent que les modèles allophoniques conduisent à un très bon compromis coût/performances (taux d'erreurs faibles avec un nombre restreint de paramètres). Néanmoins, une amélioration des performances peut être obtenue en combinant cette modélisation allophonique avec la modélisation par triphones, ceci au prix d'une augmentation du coût calculatoire.

6. BIBLIOGRAPHIE

R. Schwartz, Y. Chow, O. Kimball, S. Roucos, M. Krasner & J. Makhoul (1985), "Context-dependent modeling for acoustic-phonetic recognition of continuous speech", ICASSP, mars 1985, Tampa, Floride, USA, pp. 1205-1208.

K.F. Lee (1990), "Context-dependent phonetic hidden Markov models for continuous speech recognition", IEEE Trans. on Acoustics, Speech and Signal Processing, 1990, pp. 599-609.

L.R. Bahl, R. Bakis, J. Bellegarda, P.F. Brown, D. Burshtein, S.K. Das, P.V. de Souza, P.S. Gopalakrishnan, F. Jelinek, D. Kanevsky, R.L. Mercer, A.J. Nadas, D. Nahamoo & M.A. Picheny (1989), "Large vocabulary natural language continuous speech recognition", ICASSP, mai 1989, Glasgow, Ecosse, pp. 465-467.

K. Bartkova & D. Jouvet (1991), "Modelization of allophones in a speech recognition system", ICPhS, août 1991, Aix-en-Provence, France, Vol.4, pp. 474-477.

D. Jouvet, K. Bartkova & A. Stouff (1994), "Structure of allophonic models and reliable estimation of the contextual parameters", ICSLP, septembre 1994, Yokohama, Japon, pp.283-286.

D-DAL : UN SYSTEME DE DICTEE VOCALE DEVELOPPE SOUS L'ENVIRONNEMENT HTK

Marie-José CARATY, Claude BARRAS, Fabrice LEFÈVRE, Claude MONTACIÉ

LAFORIA-IBP – Université Paris 6 – CNRS-URA 1095 – 4 place Jussieu – 75252 Paris Cedex 05

Tél.: 44 27 47 22 – Fax: 44 27 70 00 – e-mail: caraty@laforia.ibp.fr

ABSTRACT

D-DAL is a vocal dictation system, based on Hidden Markov Models, developped under HTK (Hidden Markov model ToolKit). From the state of the art of the HMM-based systems (Gauvain & al., 1994) (Kubala & al., 1994) (Woodland & al., 1995), we have looked for the principles which are deciding for the feasability and the performance of a large vocabulary vocal dictation system. Our aim is to integrate into HTK the best technics allowing the contextual training of sub-lexical recognition units and the integration of linguistic knowledge in a word lattice. The principle and the cost of the main technics are described : the generalized contextual training, the token passing algorithm and the search of the acoustic and linguistic joint probabilities.

1. INTRODUCTION

Le but de cet article est de présenter les différents principes et techniques qui permettent l'implémentation d'un système de dictée vocale, à moyen vocabulaire (quelques dizaines de milliers de mots), fondé sur les modèles de Markov cachés. Il s'agit de décrire un système opérationnel capable de décoder le signal de parole en "phrases" (entités constituées d'une suite de mots) en utilisant au mieux le savoir-faire du décodage acoustique et les diverses sources de connaissances linguistiques. Nous décrivons ici l'essentiel des techniques qui nous ont permis de rendre opérationnel notre système D-DAL (Dictaphone-Dactylographe Automatique du Laforia) : un système de dictée vocale développé à partir de l'environnement HTK (Hidden Markov model ToolKit) (Young, 1989) (Young & al., 1992). Nous présentons les composants acoustique et linguistique de notre système.

2. COMPOSANT ACOUSTIQUE

Pour la faisabilité d'une dictée vocale à moyen vocabulaire, les unités de reconnaissance doivent être de type sub-lexical (e.g., les phonèmes). De plus, pour des raisons de performance (liée à la "maîtrise" des phénomènes de coarticulation dans la parole continue), ces unités doivent être apprises en contexte.

Sans même tenir compte de l'homophonie dans les langues (rendant indispensable l'utilisation de connaissances linguistiques) (El-Bèze, 1995) (Lamel & al., 1995), les performances d'un décodage purement acoustique sont telles qu'il est nécessaire d'utiliser un graphe de mots à partir duquel seront exploitées les connaissances linguistiques.

2.1. Modèles contextuels des unités sub-lexicales

Avec la prise en compte des contextes possibles gauche et droit, le nombre des unités devient tel qu'il est difficile de disposer de données acoustiques suffisantes pour apprendre de manière robuste la totalité des modèles. Pour contourner cette difficulté, il est possible de regrouper plusieurs contextes produisant un effet similaire sur la réalisation acoustique de l'unité sub-lexicale considérée. Le modèle appris est alors contextuel généralisé. Nous décrivons le principe d'apprentissage dans le cas particulier des phonèmes appris dans un contexte phonétique droit généralisé (biphones droits généralisés).

2.1.1. Apprentissage des biphones droits généralisés sous l'environnement HTK

HTK offre des facilités pour l'apprentissage de modèles contextuels généralisés. Nous présentons l'apprentissage des biphones droits généralisés :

- Apprentissage des modèles phonétiques indépendamment du contexte.
- Tous les biphones droits présents dans l'ensemble d'apprentissage sont dupliqués à partir des modèles phonétiques précédents.
- Ré-estimation des différents biphones droits.
- Par classe d'allophones, des liens sont établis entre les derniers états grâce à un algorithme de classification ascendante (fondé sur une mesure de divergence entre lois gaussiennes) des états finaux. Toutes les probabilités d'émission des états finaux d'un ensemble de contextes "similaires" sont ainsi liées à une seule densité de probabilité.

•Ré-estimation des différents biphones droits en tenant compte des liens établis qui réduisent le nombre de densités de probabilité à estimer et rendent plus robuste leur estimation.

Ce principe d'apprentissage est applicable au contexte gauche ou au contexte gauche-droit (Ljolje, 1994). De plus, il reste valable quelque soit l'unité sub-lexicale considérée et quelque soit l'étendue du contexte.

Les modèles de mot et de phrase sont déduits par la concaténation des modèles contextuels généralisés appropriés des unités sub-lexicales considérées. La figure 1 illustre, dans le cas d'unités phonétiques : les modèles de phonèmes (e.g., modèles à 3 états), de mots et de phrase.

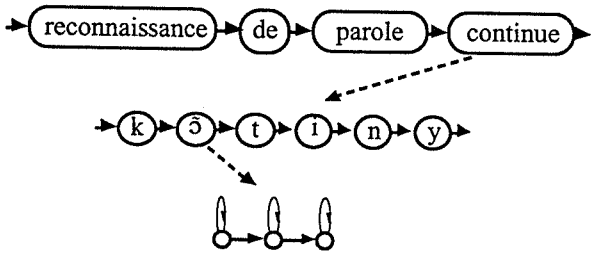


Figure 1. Modèles de phrase, de mot et de phonème.

2.2. Décodage d'une phrase à partir des modèles connectés

Le décodage d'une phrase-test à partir des modèles de mots connectés peut être effectué à partir de l'algorithme de propagation de jetons (Young & al., 1989). Cet algorithme, implémenté dans HTK, est une conception objet de la programmation dynamique qui permet d'introduire facilement et à plusieurs niveaux des contraintes sur l'enchaînement des modèles d'entités acoustiques. De plus, cet algorithme présente l'avantage d'engendrer simplement le graphe des mots décodés.

2.2.1. Algorithme de propagation de jetons

Pour faciliter la compréhension de l'algorithme, considérons l'unité sub-lexicale phonétique. Les contraintes sur l'enchaînement des modèles acoustiques sont représentées dans le modèle de propagation de jetons par un ensemble de nœuds reliés par des arcs valués. Un nœud encapsule chaque état de tous les modèles acoustiques allophoniques intervenant dans les modèles de mots. Chaque nœud contient un jeton qui représente à un instant t du décodage de la phrase-test une solution partielle du décodage (à cet instant). En terme de chemin, cette solution partielle est constituée du sous-chemin décrivant la suite des mots déjà décodés et du chemin (dans le modèle de mot) arrivant à l'état markovien (du modèle allophonique) associé au nœud considéré. La valeur de l'arc représente une densité de probabilité d'émission (mise à jour au décodage de chaque trame de test) à laquelle on peut ajouter une valeur fixe dépendant des contraintes linguistiques (e.g., le modèle des bigrammes). Parmi l'ensemble des nœuds, on étiquette ceux qui correspondent aux fins de modèles de mots.

Au décodage de chaque trame de test : pour tout nœud, on propage son jeton vers les nœuds suivants en incrémentant sa valeur de la valuation de l'arc emprunté. A partir de l'ensemble des jetons arrivant sur un nœud, on attribue un nouveau jeton au nœud (e.g., le meilleur jeton au sens probabiliste). Si le nœud est étiqueté comme fin de mot, on ajoute ce mot au sous-chemin optimal associé au jeton. On construit alors le graphe de mots en mémorisant, à chaque instant t du décodage, les N meilleurs jetons de fin de mot représentant les N meilleurs sous-chemins.

La complexité de cet algorithme est donc égale au coût de traitement d'une trame multiplié par le nombre de trames. Le coût de traitement d'une trame se décompose en une partie acoustique et une partie linguistique. Soit N_{al} le nombre d'allophones et N_m le nombre de mots du vocabulaire : la complexité de la partie acoustique est en $O(N_{al})$, la complexité de la partie linguistique dépend du type de contraintes utilisées. Elle est en $O(N_m)$ dans le cas où il n'y a pas de contrainte et en $O(N_m^2)$ si l'on utilise un bigramme de mots.

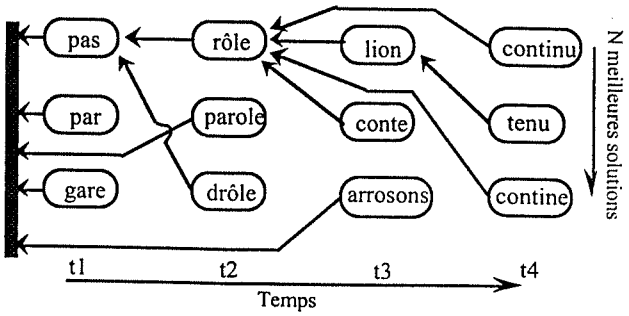


Figure 2 : Exemple de graphe de mots.

3. COMPOSANT LINGUISTIQUE

Dans le cadre markovien, l'utilisation de modèles stochastiques de langage est naturelle. En effet, la succession des mots dans une phrase est soumise à des contraintes grammaticales et sémantiques. Le principe est alors d'estimer la probabilité de l'occurrence d'un mot, connaissant le début de la phrase. C'est ce type de connaissance linguistique qui est utilisé pour améliorer le décodage acoustique.

3.1. Modèle stochastique de langage

La probabilité d'une suite de k mots $m_1 ... m_k$ est exprimée comme le produit des probabilités conditionnelles suivantes:

$$P(m_1,...,m_k) = P(m_1)\prod_{i=2}^{k} P(m_i/m_1,...,m_{i-1})$$

Pour des raisons de faisabilité, il est nécessaire d'apporter des simplifications à ce modèle. Les modèles k-grammes de langage sont ainsi caractérisés par le fait que la probabilité d'apparition d'un mot ne dépend que des *k-1* mots précédents. Les modèles k-classes sont une généralisation des modèles k-grammes, pour laquelle on ne considère plus le mot mais sa caractérisation grammaticale.

3.1.1. Apprentissage des k-grammes

Les corpus écrits que nous utilisons pour l'apprentissage des k-grammes proviennent du projet Gutenberg pour l'anglais et d'ABU pour le français. Les textes sont essentiellement des textes littéraires.

Si l'on note $C(m_1...m_k)$ le nombre d'apparitions de la suite de mots $m_1...m_k$ dans le corpus d'apprentissage, un modèle k-grammes est estimé par la formule :

$$\tilde{P}(m_k / m_1,...,m_{k-1}) = \frac{C(m_1...m_k)}{C(m_1...m_{k-1})} = \tilde{P}_k$$

La taille des corpus écrits est généralement insuffisante pour une mesure fiable des fréquences de toutes les suites de k mots pour k=3 ou plus. Par conséquent, on préfère estimer les modèles k-grammes (pour k≥3) par une interpolation des j-grammes (j=1,...,k-1). Ce qui donne pour un trigramme :

$$\hat{P}(m_3 / m_1, m_2) = \lambda_1^{12}\tilde{P}_1 + \lambda_2^{12}\tilde{P}_2 + \lambda_3^{12}\tilde{P}_3$$

où les jeux de facteurs $\{\lambda_i^{12}\}$ (i=1,...3) sont ajustés automatiquement en fonction du nombre d'occurrences du couple m_1m_2 de manière à rendre les données observées plus probables selon le critère du maximum de vraisemblance (Bahl & al., 1983) (Jelinek, 1989).

Le principe est de trouver le jeu de paramètres qui maximise la probabilité du trigramme $m_1m_2m_3$ quelque soit m_3. Pour ce faire, on prend un nouveau texte sur lequel on calcule pour chaque couple m_1m_2 la contribution de chacun des k-grammes (k=1,...,3) au calcul d'un trigramme $m_1m_2m_3$. A partir de ces contributions, on peut calculer un nouveau jeu de paramètres et donc en obtenir une bonne estimation par itérations successives.

3.2. Intégration des modèles de langage

L'intérêt des modèles de langage est évidemment de pouvoir rétablir, du moins on peut l'espérer, la "meilleure" suite de mots au sens de l'utilisation conjointe des connaissances acoustique et linguistique. La solution est donnée par un algorithme de programmation dynamique fondé sur la connaissance acoustique (le graphe de mots décodés) et sur la connaissance des k-grammes du langage.

3.2.1. Algorithme d'intégration des probabilités conjointes

L'algorithme décrit ici pour les trigrammes du langage est généralisable aux k-grammes. Le graphe de mots est composé de N mots par trame. Les mots ont été mémorisés sur leur trame finale et possèdent un pointeur sur leur trame initiale (figure 2). A l'instant t, pour chaque mot on peut associer les N "meilleurs" mots l'ayant précédé. Pour chaque trame, on construit ainsi un tableau de N^2 couples de mots.

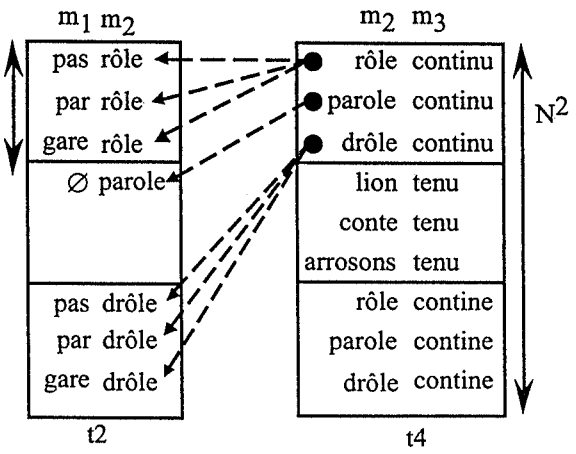


Figure 3: Illustration de l'algorithme de calcul des probabilités conjointes.

La figure 3 donne les tableaux ainsi construits pour les trames correspondant aux temps t_2 et t_4 de l'exemple de graphe de mots illustré à la figure 2. Considérons le tableau des couples de mots m_2m_3 (temps t_4), pour chaque couple m_2m_3 on estime le meilleur mot précédent parmi les N mots m_1 possibles avant m_2 (figure 3). Le "meilleur" mot est obtenu par comparaison des probabilités conjointes acoustique-modèle de langage calculée comme le produit de la probabilité conjointe jusqu'à m_2 par la probabilité acoustique associée à m_3 et par la probabilité du trigramme $m_1m_2m_3$. Un lien est alors créé du couple m_2m_3 vers le couple m_1m_2 On itère ce processus de la première à la dernière trame par programmation dynamique. La meilleure phrase décodée est

alors reconstruite à partir du couple de la dernière trame ayant la plus forte probabilité conjointe. La complexité de cet algorithme est en $O(N^3)$.

5. CONCLUSION

Le système D-DAL est construit à partir d'un environnement largement répandu (HTK) et dans lequel on peut intégrer les dernières recherches, que ce soit au niveau de la modélisation acoustique ou de la modélisation linguistique. Ce système est proche de ce que l'on peut attendre d'un système de référence (Chollet & al., 1982) (El-bèze, 1995).

D-DAL sera l'un des systèmes participant à la première campagne de tests, sur la dictée vocale en langue française, financée par l'AUPELF-UREF.

6. BIBLIOGRAPHIE

Bahl L., Jelinek F., Mercer R. (1983) "A Maximum Likelihood Approach to Continuous Speech Recognition", IEEE Trans. on PAMI, PAMI-5 vol.2

Barras C. (1996) "Reconnaissance de la parole continue: Addaptation au locuteur et contrôle temporel dans les modèles de Markov cachés", Thèse de l'Université Paris 6 (à paraître)

Chollet G., Gagnoulet C. (1982) "On the Evaluation of Speech Recognizers and Data Bases using a Reference System", ICASSP, 2026-2029

El-Bèze M. (1995) "Utilisation des modèles stochastiques de langage", Ecole thématique-Fondements et Perspectives en Traitement Automatique de la Parole, Ed. H. Méloni, 129-138

Gauvain J.-L., Lamel L., Adda G., Adda-Decker M. (1994) "Speaker-independant Speech Dictation", Speech Communications, vol. 15, 21-37

Jelinek F. (1989) "Self-Organized Language Modeling for Speech Recognition", Readings in Speech Recognition, Morgan Kaufman

Kubala F., Anastasakos A., Makhoul J., Nguyen L., Schwartz R., Zavaliagkos G. (1994) "Comparative experiments on Large Vocabulary Speech Recognition", ICASSP, vol. 1, 561-564

Lamel L., Adda-Decker M., Adda G., Gauvain J.-L. (1995) "Reconnaissance Multilingue de Grands Vocabulaires", Ecole thématique-Fondements et Perspectives en Traitement Automatique de la Parole, Ed. H. Méloni, 119-128

Ljolje A. (1994) "High Accuracy Phone Recognition using Context Clustering and Quasi-Triphonic Models", Computer Speech and Language, vol. 8, 129-151

Woodland P.C., Leggetter C.J., Odell J.J., Valtchev V., Young S.J. (1995) "The 1994 HTK Large Vocabulary Speech Recognition System", ICASSP, 73-76

Young S.J., Russel N.H., Thornton J.H.S. (1989) "Token Passing : a Simple Conceptual Model for Connected Speech Recognition Systems", Technical Report CUED/F-INFENG/TR38, Cambridge University Engineering Dpt.

Young S.J. (1992) "HTK Version 1.4: Reference Manual and User Manual", Cambridge University Engineering Department - Speech Group

ANNEXE

Nous donnons ici un exemple de dictée vocale. Nous avons choisi de décoder 192 phrases de la base de données TIMIT. Ces phrases, comportant 1570 mots, sont prononcées par 24 locuteurs. Les modèles acoustiques sont les 404 biphones droits généralisés appris sur 3696 phrases prononcées par 462 locuteurs (Barras, 1996). Le lexique comporte 6107 mots dont la phonétisation normative est fournie. Dans un but d'illustration, nous avons choisi, comme modèle de langage, un modèle des bigrammes et des trigrammes de l'ensemble des 2342 phrases de TIMIT (dont les phrases de test). Le facteur de branchement est alors de 115. Nous avons choisi, comme profondeur (N) du graphe de mots, 1 % de la taille du lexique soit N = 60.

Quatre exemples de décodage d'une phrase prononcée (P) sont donnés. Le premier est le décodage acoustique pur (D1), le deuxième correspond au post-traitement par le modèle des trigrammes du premier décodage (D2), le troisième intègre le modèle des bigrammes dans le premier décodage (D3), le quatrième combine les modèles des bigrammes et des trigrammes (D4).

P) Will Robin wear a yellow lily.
D1) Wool Robin worries ya elm lily.
D2) Will Robin worried elbowing
D3) Will Robin were a yellow lily.
D4) Will Robin wear a yellow lily.

Table 1: Résultats de reconnaissance : nombre de mots identifiés (Id.), substitués (Subs.), omis (Omis), insérés (Ins.) et pourcentage de reconnaissance (% Rec.).

	Id.	Subs.	Omis	Ins.	% Rec.
D1	481	1044	45	449	**2 %**
D2	955	467	148	59	**57.1 %**
D3	1332	178	60	46	**81.9 %**
D4	1439	77	54	11	**91 %**

DÉCODAGE ACOUSTICO-PHONÉTIQUE FLOU

Olivier OPPIZZI, David FOURNIER, Philippe GILLES, Henri MÉLONI
CERI - Laboratoire d'informatique, 339 chemin des Meinajariès - BP 1228 - 84911 AVIGNON Cedex 9
Tel: 90 84 35 18 - Fax: 90 84 35 01 - e-mail: oppizzi@univ-avignon

ABSTRACT

In this paper, a general framework of acoustic-phonetic modelling is developed. Context sensitive rules are incorporated and assessed into a speech recognition system based on fuzzy decision making. Using fuzzy sets, the platform provides a reliability measure in order to gain knowledge about the behaviour of each rule and to perform rational fusion operators. A particular attention is paid to rules integration: although values returned by rules can be either of numerical or of symbolic nature, they have been adjusted to fit the dynamic ranges of the acoustic cues.

The results on a test set of isolated words with a one-speaker low training rate are carried out via a two steps procedure: a one-speaker speech database and a test bed with 4 speakers who were never involved during the training. Further investigations are needed to improve the performance.

1. INTRODUCTION

Qu'il s'agisse d'élaborer un modèle mathématique qui rende compte de la solution optimale ou qu'il s'agisse d'expérimenter des heuristiques *ad hoc*, les mécanismes de décision tiennent une place prépondérante dans les systèmes informatiques pour lesquels il faut choisir une solution parmi d'autres.

Approche possible au problème de la décision (Dubois, 1983, chap.6) initiée par (Zadeh, 1965), les ensembles flous ont la caractéristique d'être un modèle intuitif de représentation de l'imprécis et de l'incertain: un ensemble contient un élément à un degré plus ou moins élevé, ce degré étant établi par une fonction d'appartenance.

La présente recherche permet d'expérimenter la décision floue au sein d'un système de décodage acoustico-phonétique (DAP) à base de règles contextuelles. Plusieurs unités phonétiques pouvant être détectées sur un segment temporel s donné, la décision consiste dans un premier temps à attribuer au couple <règle,unité phonétique> une mesure de confiance calculée à partir du modèle des ensembles flous. Dans un deuxième temps, afin de rendre possible le choix de la "meilleure" unité sur s, les mesures de confiance seront agrégées en une note pour l'unité, toutes règles confondues, puis en une note pour le mot.

2. LE DAP DESCENDANT

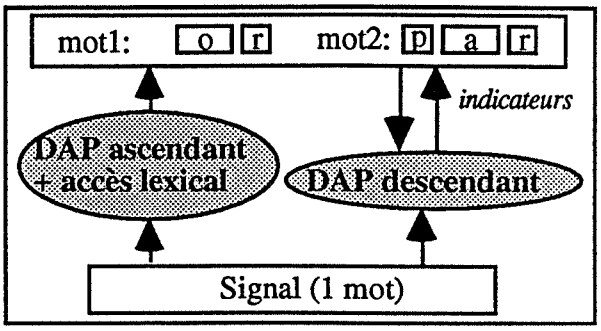


Figure 1: Principe du DAP.

Le système de reconnaissance de mots isolés que nous avons utilisé a été exploité dans sa phase descendante (figure 1).

Définition: unité phonétique.

Une unité phonétique est un couple <p, s> où p est le phonème et s le segment sur lequel p a été détecté par le DAP ascendant.

Définition: Unité contextuelle (triphone).

Il s'agit d'un triplet $<u_g, u_c, u_d>$. u_c est une unité phonétique sur un segment s, u_g l'unité immédiatement à gauche et u_d l'unité immédiatement à droite.

Définition: indicateur.

Un indicateur i est une valeur produite par une règle R de décodage acoustico-phonétique appliquée à une unité contextuelle u: i=R(u). Les indicateurs sont hétérogènes entre règles distinctes (une règle peut produire des symboles, des scores, des indices acoustiques etc.).

La phase ascendante (Gilles, 1993) a pour rôle de segmenter le signal et d'identifier les unités phonétiques sans vérification d'indices acoustiques de coarticulation. Dans un second temps, un module lexical (Béchet, 1994) produit une cohorte de mots candidats. Enfin, connaissant dans le treillis le contexte phonétique et segmental proche de chaque hypothèse phonétique, il devient possible de vérifier des indices acoustiques contextuels en phase descendante.

Exemple de règle contextuelle

Les voyelles fermées s'ouvrent en contexte /r/ (F1 croît). La bande de

fréquence où la règle recherche F1 est élargie.

Comment décider d'infirmer ou confirmer une hypothèse phonétique donnée au su des informations apportées par les règles descendantes ?

Initialement, les règles seuillaient les indices calculés puis s'inscrivaient dans une architecture de décision hiérarchique. Le contrôle réalisé par la règle `regle_FFSSCH()` pour la détection des fricatives sourdes est illustré (figure 2). La règle `regle_FF()` discrimine le /f/ des deux autres fricatives sourdes en vérifiant que l'unité en entrée est diffuse (faible variation de l'amplitude spectrale). La règle `regle_SSCH()` discrimine le /s/ du /ʃ/ en contrôlant l'existence d'une pente spectrale en basses fréquences pour /ʃ/ et en hautes fréquences pour /s/.

```
règle FF(u=<pho,seg>)
    S = spectre de la trame stable de seg
    Max = amplitude max de S en hautes fréquences
    Min = amplitude min de S en basses fréquences
    si( |Max-Min| < Seuil ) alors renvoi oui (c'est un /f/)
règle SSCH(u=<pho,seg>)
    t = trame de seg où le spectre varie le plus
    S = spectre en trame t
    F = fréquence de S où la dérivée est maximale
    si( F < Seuil ) alors renvoi /ʃ/, sinon renvoi /s/
CONTROLE: règle FFSSCH(u)
si( regle_FF(u) rend oui ) alors renvoi /f/
|_                          sinon renvoi regle_SSCH(u)
```

Figure 2: Décision hiérarchique avec seuil.

Une telle prise de décision est dangereuse: il y a risque, en fondant la décision sur un seuil, qu'une faible variation des phénomènes acoustiques engendre une réaction opposée du système (Dubois, 1983, p.156).

Par ailleurs, il existe des règles concurrentes susceptibles de fournir des résultats contradictoires pour la détection d'une même unité contextuelle donnée. La mesure d'incertitude introduite en section 3, en permettant d'apprendre les capacités de reconnaissance des règles, offre la possibilité de gérer rationnellement les conflits.

3. DÉCISION FLOUE

Les modèles flous ne sont pas étrangers aux systèmes de reconnaissance de la parole. (De Mori, 1983, p.109) décrit l'incertitude d'indices acoustiques par des ensembles flous (figure 3) et applique la fusion bayésienne sur ces valeurs d'incertitude pour obtenir l'incertitude relative au trait "voisé".

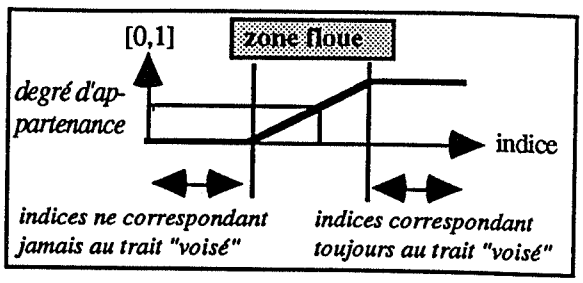


Figure 3: Ensemble flou d'indices acoustiques.

De part la forte variabilité des indices acoustiques de coarticulation, les règles que nous utilisons ne permettent pas de localiser aussi simplement la zone floue. Par ailleurs, nous avons apporté plus de nuances dans la description de l'incertitude qu'une simple interpolation linéaire entre 0 et 1 en conduisant un apprentissage automatique des capacités de reconnaissance de chaque règle (figure 4).

En phase de reconnaissance, une fois appris le comportement de chaque règle, une mesure de confiance $C_R(u)$ à valeurs dans [0,1] est alors attribuée à un triphone u sachant que la règle R a produit l'indicateur `i=R(u)`.

Plus largement détaillée dans (Oppizzi, 1995, chap.5) (figure 4), cette mesure est applicable à des indicateurs de nature quelconque (règles renvoyant des valeurs binaires, probabilistes, symboliques); elle est uniquement contrainte par la nécessité de disposer d'une relation d'ordre qui soit définie sur le domaine de définition des indicateurs de chaque règle.

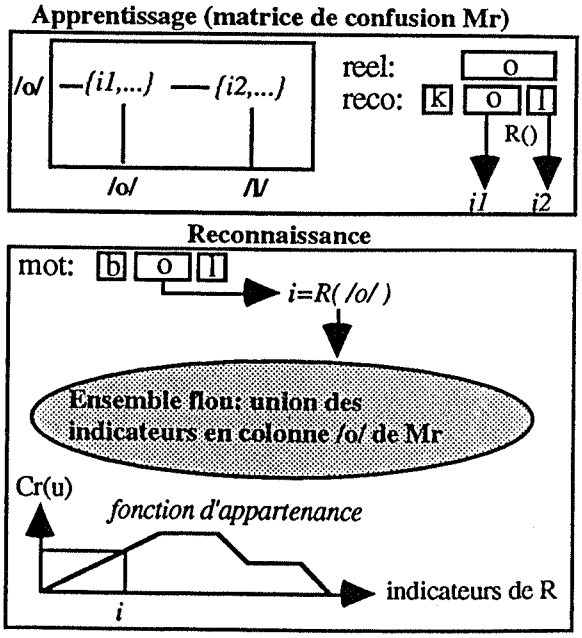


Figure 4: Calcul de la mesure de confiance.

Définition: confusion.

Une règle R confond deux triphones u_1 et u_2 distincts si $R(u_1)=R(u_2)$.

La mesure de confiance permet d'exprimer un degré d'incertitude correspondant à un degré de confusion. En phase d'apprentissage, chaque règle R est appliquée à chaque phonème reconnu et la valeur obtenue est ventilée dans une matrice de confusion M_R (figure 4). Chaque entrée de la matrice correspond donc à une liste d'indicateurs. Les phonèmes réels du corpus apparaissent en ligne, les unités reconnues u en colonne; sur l'exemple, "eau" correspond au mot réellement prononcé, "colle" au mot reconnu par le DAP ascendant.

Pendant la reconnaissance, pour calculer la confiance $C_R(u)$, les indicateurs en colonne u de M_R sont regroupés dans un ensemble flou et la fonction d'appartenance tracée. Si R est une règle markovienne, les indicateurs regroupés dans la matrice de confusion et dans l'ensemble flou sont des probabilités. Ces probabilités sont installées sur l'axe des abscisses de la fonction d'appartenance au moyen de la relation d'ordre sur les nombres, et la valeur de la mesure de confiance est lue en ordonnée. L'interprétation de cette mesure est reprise en illustration (table 1) pour des valeurs extrêmes.

Table 1: Interprétation de $C_R(u)$

$C_R(u)$	**Interprétation**
1	Certitude que u est présente dans le signal de parole. R ne la confond pas, et i=R(u) a été appris.
0,5	Incertitude sur la présence de u dans le signal de parole. R la confond.
0	Certitude que u n'est pas présente dans le signal. R ne la confond pas, mais i=R(u'), u'≠u, a été appris.

Le contrôle flou entre règles R_i correspond à un contrôle sans hiérarchie (figure 5).

mot 'vil':	V	I	L
$C_{R1}()$:	c_{11}	c_{12}	c_{13}
$C_{R2}()$:	c_{21}	c_{22}	c_{23}
•••			
fusion 2 <- (fusion 1	fusion 1	fusion 1)	

Figure 5: Décision floue.

Le module lexical relève dans le treillis des chemins d'unités phonétiques correspondant à des mots. Chaque règle R_i est alors appliquée à chaque unité contextuelle de chaque mot avant que la mesure de confiance $C_{Ri}()$ ne soit appliquée aux valeurs produites par ces règles.

L'opérateur fusion 1 en colonne permet d'obtenir une unique mesure de confiance par unité phonétique, ces mesures pouvant être ensuite agrégées par l'opérateur fusion 2 afin d'attribuer une confiance au mot.

4. RÉÉCRITURE DES RÈGLES

4.1. Indicateurs fournis par les règles

Les règles consistent à détecter divers indices acoustiques qu'il faut alors agréger pour produire un indicateur de sortie non booléen.

Exemple: détection des voyelles aiguës.

La règle évalue les densités d'énergie spectrale D_a, D_c et D_g dans les bandes de fréquences où F2 est attendu respectivement pour les voyelles aiguës, centrales et graves. Différentes fonctions d'agrégation sont analysées (table 2) afin d'établir quel indicateur i la règle doit renvoyer. '+' et '-' indiquent le signe du résultat attendu de la fusion. Le deuxième opérateur est celui qui *a priori* discrimine au mieux les bons cas (voyelle aiguë, donc D_a est majoré) des mauvais cas de reconnaissance (voyelle centrale ou grave, donc D_a est minoré).

Table 2: Agrégation pour voyelles aiguës.

	voy. aiguë	*autres voy.*
$D_a - D_c - D_g$	-	-
$D_a - \max(D_c, D_g)$	+	-
$D_a - \min(D_c, D_g)$	+	0

4.2. Robustesse

En décision floue, les règles ne sont pas appliquées hiérarchiquement mais "à plat" en tant que règles concurrentes. Il n'est donc pas possible de contrôler acoustiquement les triphones avant de les envoyer aux règles.

Exemple: les fricatives sourdes.

La règle `regle_SSCH()` (figure 2) ne signifie plus rien si elle est appliquée en détection de /f/. Appliquées à des unités non prévues, elle peut générer par hasard un bon résultat de reconnaissance. Cette règle a donc été réécrite de façon à ce que sa robustesse soit augmentée: elle accepte en entrée et est capable de discriminer une fricative sourde quelconque.

5. RÉSULTATS

Trente-trois règles descendantes dont dix règles contextuelles ont été programmées et testées en reconnaissance de mots isolés. Un apprentissage volontairement restreint des performances du DAP ascendant a été mené — 85 mots choisis dans BDLEX et prononcés par un seul locuteur. Les règles ont été appliquées à chaque unité contextuelle du corpus et les résultats obtenus distribués dans la matrice de confusion.

Relativement à un vocabulaire de 18000 mots, le premier jeu de test (figure 6) illustre la reconnaissance de 800 mots prononcés par le locuteur sur lequel l'apprentissage a eu lieu. Le classement du bon mot reconnu au sein de la cohorte de mots candidats se lit en abscisse, le pourcentage cumulé de bonne reconnaissance en ordonnée (90% des mots bien reconnus sont apparus dans les 23 premiers mots de leur cohorte).

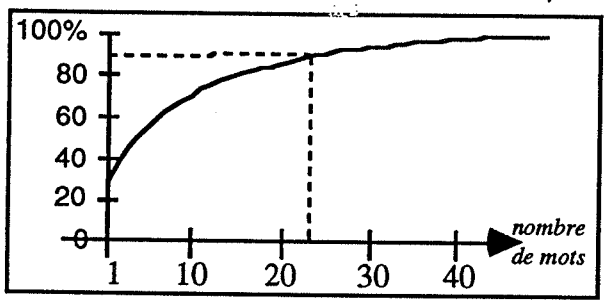


Figure 6: Classement du bon mot, 1 locuteur.

Le second jeu de test (figure 7 et 8) a été réalisé pour 4 locuteurs autres que le locuteur pour lequel l'apprentissage a eu lieu, mais du même sexe. Ainsi, le test rend compte de l'indépendance des règles au locuteur.

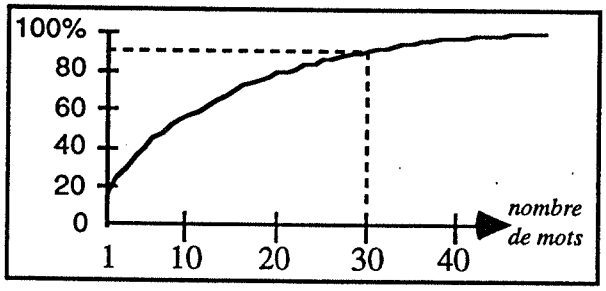


Figure 7: Classement du bon mot, 4 locuteurs.

Ce résultat est complété (figure 8) par une évaluation de l'écart entre les classements lexicaux proposés par le DAP ascendant et le système de décision floue. L'abscisse positive correspond à des mots mieux classés dans les cohortes après décision floue.

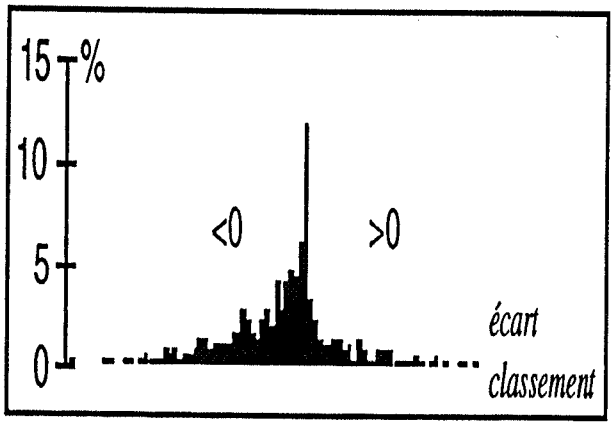


Figure 8: Gain en décision floue, 4 locuteurs.

La dépréciation observée des résultats s'explique de plusieurs façons:

- Nombre de règles

  De nombreux triphones ne subissent pas d'analyse contextuelle: le nombre et la variété des règles ne sont pas suffisants.

- Carences de l'apprentissage

  La mesure de confiance est sujette à caution. L'apprentissage étant limité, le module de décision se montre capable d'attribuer une forte confiance $C_R(u)$ même si peu de tests ont été recueillis pour le couple $<R,u>$.

6. PERSPECTIVES

Un plus grand nombre de règles contextuelles est nécessaire pour améliorer la reconnaissance des triphones. Le système permettant d'évaluer le comportement des règles existantes, leur potentiel discriminant peut être augmenté.

D'autre part, le couplage d'un système orienté connaissances à des systèmes markoviens et neuro-mimétiques pourrait permettre de valider le calcul de confiance sur des indicateurs probabilistes ou vectoriels.

L'effort à venir concerne enfin la justification théorique des valeurs utilisées pour la paramétrisation du système, notamment dans le sens d'établir l'optimalité du choix des fonctions d'agrégation.

7. BIBLIOGRAPHIE

Béchet F. (1994) *Système de traitement de connaissances phonétiques et lexicales: application à la reconnaissance de mots isolés sur de grands vocabulaires et à la recherche de mots cibles dans un discours continu*, Thèse de l'Université d'Avignon

De Mori R. (1983) *Computer Models of Speech using Fuzzy Algorithms*, Plenum Press, New York

Dubois D. (1983) *Modèles mathématiques de l'imprécis et de l'incertain en vue d'applications aux techniques d'aide à la décision*, Thèse de l'Institut National Polytechnique de Grenoble

Gilles P. (1993) *Décodage phonétique de la parole et adaptation au locuteur*, Thèse de l'Université d'Avignon

Oppizzi O., Méloni H. (1995) Comparer l'incomparable, *Actes des Rencontres Francophones sur la Logique Floue et ses Applications (LFA)*, 256-264

Zadeh L.A. (1965) Fuzzy Sets, *Information Control*, n°8, 338-353

$ETC_{vérif}$ UN ENVIRONNEMENT MULTI-AGENTS DE RECONNAISSANCE AUTOMATIQUE DE LA PAROLE EN CONTINU

Jean-Luc COCHARD Murielle VIAL

IDIAP, Case postale 592, CH-1920 Martigny

Tél: (41) 26 22 76 64 - Fax: (41) 26 22 78 18 - e-mail: [cochard vial]@idiap.ch

Abstract

$ETC_{vérif}$ is a prototype of a cooperative automatic speech recognition system. This work stems from the strong intuition that a probable solution to the general problem of speech understanding lies in the development of a system able to deal with a large set of distinct, partial and even unreliable problem solvers, namely HMMs (Hidden Markov Models), GTP (Graphemes To Phonemes) agents, prosodic analysers and even higher order agents processing syntactic and semantic knowledge. Up to now, these solvers allow us to work with words and phonemes. This paper describes how a composite solution is computed from all these partial solutions.

1 INTRODUCTION

$ETC_{vérif}$ [CO95, CF95] est un système multi-agents, basé sur la plate-forme ETC (Environnement de Traitement Coopératif) qui fournit différents services:

1. Un modèle d'Ω-*agents* constituant les différentes sources de connaissances du système;
2. Un modèle de μ-*agents* qui sont les hypothèses fournies par les Ω-*agents*;
3. Un moteur d'évolution du système.

Ce projet, réalisé conjointement à un autre projet de collection de plusieurs bases de données du français parlé en Suisse romande, sert de vérification automatique partielle de la cohérence du contenu de bases de données.

Ses données d'entrée sont de deux sortes : le signal de parole et le texte que doit prononcer le locuteur. La sortie indique si c'est effectivement ce texte qui a été dit.

2 ARCHITECTURE DU SYSTEME

Le prototype $ETC_{vérif}$ s'inspire des travaux effectués par Susan E. Lander [Lan94], proposant notamment une architecture modulaire et flexible dans laquelle la résolution de problèmes est basée sur un ensemble d'agents (voir aussi [HCA94]) et la recherche d'une solution globale est orientée par un protocole de négociation générique. La spécificité de notre système est de contenir deux niveaux d'agents (voir figure 1 et [CO95]).

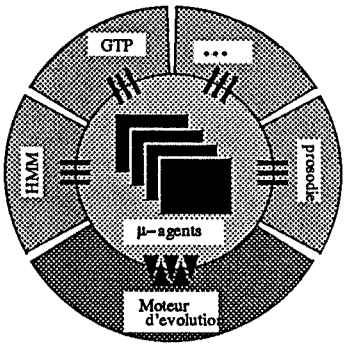


Figure 1: Architecture de $ETC_{vérif}$.

2.1 Les Ω-agents

Les Ω-*agents* ont chacun leur domaine de connaissance et apportent une solution locale au problème de la reconnaissance de la parole.
Deux fonctions leur sont attribuées:

1. Initialisation d'une solution: l'agent accepte en entrée la donnée complète d'un problème qu'il sait traiter localement et il transmet en sortie une solution satisfaisant au mieux ses contraintes locales;
2. Révision localisée: L'agent accepte en entrée une référence à une donnée sur laquelle il

a déjà fourni une solution initiale. Cette référence est assortie d'une indication de la zone de la donnée concernée par cette demande. L'agent transmet en sortie une nouvelle interprétation de cette zone en tenant compte des solutions qu'il a déjà fournies.

Les Ω-*agents*, actuellement utilisés dans notre système, sont des HMMs (Hidden Markov Models) [Tor94] et un agent GTP (Graphemes To Phonemes). L'agent GTP fournit une séquence de phonèmes à partir d'une phrase en codes ASCII. Cette séquence est ensuite utilisée comme grammaire de l'agent HMM basé sur des modèles de phonèmes afin de connaître la position temporelle de ces derniers. Deux agents HMM sont disponibles: l'un produit des phonèmes, l'autre des mots. Ce dernier peut également fournir des phonèmes en collaborant avec l'agent GTP: son entrée est le signal de parole, une sortie intermédiaire est une séquence de mots, cette séquence devient l'entrée de l'agent GTP et le résultat final est une transcription phonétique de la séquence de mots qui deviendra, une nouvelle fois, la grammaire de l'agent HMM basé sur des modèles de phonèmes.

Ainsi, le système dispose de trois séquences de phonèmes obtenues par les Ω-*agents* GTP, HMM s'appuyant sur des modèles de phonèmes et HMM s'appuyant sur des modèles de mots combinés avec l'agent GTP. Par ailleurs, il dispose aussi de deux séquences de mots qui sont le texte devant être prononcé par le locuteur (i.e. l'entrée de l'agent GTP) et le résultat fourni par l'agent HMM s'appuyant sur des modèles de mots.

2.2 Les μ-agents

Les *μ-agents* sont les résultats fournis par les Ω-*agents*. Ils peut s'agir de phonèmes ou de mots. Cinq informations sont stockées dans ces agents:

1. Borne_inf: borne temporelle inférieure;
2. Borne_sup: borne temporelle supérieure;
3. Id: l'identificateur du *μ-agent* (par exemple, une étiquette phonétique);
4. Les liens de voisinage
5. La confiance locale

2.3 Fonctionnement

Les Ω-*agents* produisent une première solution décomposée en une séquence de *μ-agents* qui sont des mots et des phonèmes. Ces derniers doivent alors estimer leur fiabilité (voir section 3.2). Le "moteur d'évolution" sélectionne ensuite la meilleure solution courante, c'est-à-dire la séquence de *μ-agents* situés sur le chemin optimal (voir section 3.3). Enfin, ce chemin est parcouru afin de trouver les zones temporelles de faible fiabilité dans lesquelles les Ω-*agents* devront fournir de nouvelles solutions (voir section 3.4).

3 IMPLEMENTATION

Les différentes étapes de calcul, détaillées ci-dessous, interviennent dans un système dynamique dont la stratégie de résolution est menée par des événements: l'adjonction de *μ-agents* ou la modification de leurs informations locales.

3.1 Initialisation d'une solution

Les Ω-*agents* sont des entités qui attendent une entrée spécifique afin de produire une première solution. Celle-ci est fournie sous forme d'une séquence de *μ-agents* (phonèmes, mots) qui sont des entités autonomes capables de déterminer leur fiabilité en se mettant en correspondance les unes avec les autre. Ce facteur détermine le taux d'intégration d'une hypothèse codée par un *μ-agent* dans son environnement.

Les relations établies entre deux *μ-agents* sont de deux types:

1. Les relations horizontales – Ce type de relation s'établit entre *μ-agents* de même classe, c'est-à-dire qui contiennent des informations comparables. Ainsi, les relations suivantes peuvent être établies entre deux *μ-agents* A et B:
   - A préc B: A a une valeur de borne temporelle supérieure qui coïncide avec la borne inférieure de B;
   - A suit B: A a une valeur de borne temporelle inférieure qui coïncide avec la borne supérieure de B;
   - A équiv B: les bornes temporelles inférieures et supérieures de A et B sont très semblables et leurs étiquettes E(A) et E(B) font partie d'une même classe phonétique. A et B sont alors des *μ-agents frères*.

   Quand l'une de ces propriétés est vérifiée, deux *μ-agents* deviennent voisins.
2. Les relations verticales – Ce type de relation décrit les liens entre *μ-agents* de classes différentes, par exemple, entre un phonème et un mot. En effet, chaque mot connaît sa décomposition phonétique grâce à l'agent GTP et est

capable de savoir si une suite de phonèmes du système y correspond.

Les calculs suivants (fiabilité et recherche du chemin optimal) ne sont effectués que par les phonèmes. En effet, d'abord il ne serait pas très judicieux de les faire pour les mots car on ne dispose que de deux séquences de ceux-ci (voir section 2.1): par conséquent, s'il existe deux mots différents sur une même zone temporelle, aucun indice ne peut indiquer lequel est le plus fiable. Ensuite les phonèmes (sauf ceux générés par l'Ω-*agent* HMM s'appuyant sur des modèles de phonèmes) sont obtenus par phonétisation (grâce à l'Ω-*agent* GTP) de ces mots. Ainsi, si un mot a une forte fiabilité, cela a une répercussion directe sur les phonèmes. De ce fait, les mots vont servir à renforcer et améliorer la solution finale, composée de phonèmes. Si dans une certaine zone temporelle, une suite de phonèmes, contenue dans cette solution, correspond partiellement à un mot proposé par le système à l'intérieur de cette zone, il sera possible de rechercher les phonèmes manquants et, s'ils ont une fiabilité suffisamment élevée, de les intégrer dans la solution finale. Le système sera alors en mesure de donner la séquence de mots qui a été prononcée dans les zones de forte fiabilité tandis que, dans celles de faible fiabilité, les Ω-*agents* devront fournir de nouvelles solutions et recommencer à négocier.

3.2 Calcul de la fiabilité d'un μ-agent

La fiabilité des μ-*agents* dépend de leur niveau d'intégration dans leur contexte. Afin de déterminer cette mesure, ils vont se communiquer différents résultats plutôt que de se voir attribuer des scores, de manière plus classique mais sans échange d'informations.

Les relations (horizontales) établies précédemment sont utilisées pour déterminer la fiabilité des μ-*agents*. Ainsi, ces derniers sont d'autant plus fiables que leurs voisins sont nombreux et ont, eux-mêmes, une fiabilité élevée. Cette mesure est dynamique et évolue avec l'adjonction de nouveaux μ-*agents* dans le système et la création de nouvelles relations.

Le calcul de confiance est itératif. Soit $F_i(A)$ la fiabilité du μ-*agent* A à l'itération i. On a les relations suivantes:

$$\begin{aligned} F_i(A) &= F_{i-1}(A) - \frac{1}{(2+Fv_{i-1}(A))^i} \\ F_0(A) &= 1 \end{aligned}$$

$Fv_{i-1}(A)$ est la somme des fiabilités des voisins de A à l'itération $i-1$. Le calcul de la fiabilité de A se termine lorsque $F_i(A)$ se stabilise, c'est-à-dire lorsque:

$$F_{i-1}(A) - F_i(A) < \Delta$$
$$\Delta = 0.01$$

Cette valeur permet de parvenir à une bonne approximation de la valeur stable sans avoir des temps de calcul trop longs. De plus, la définition de $F_i(A)$ garantit que la confiance converge dans l'intervalle $]0,1[$ puisque $Fv_i(A) > 0$.

3.3 Recherche du chemin optimal

Le chemin optimal est la séquence de phonèmes de plus forte fiabilité. La méthode la plus simple est de choisir celui sur lequel la somme des fiabilités est la plus forte mais comme le montre la figure 2, on choisit alors le chemin comportant le plus grand nombre de phonèmes (dans notre exemple, B, C et D plutôt que A et D car 0.50 + 0.60 + 0.65 > 0.70 + 0.65), ce qui n'est pas le but recherché.

Aussi, choisit-on de calculer la somme des fiabilités pondérées par les durées des μ-*agents* afin que le chemin optimal comporte la séquence la plus fiable. Ainsi, dans la figure 2, si on considère que les larges rectangles ont une durée deux fois plus élevée que les rectangles étroits, le chemin optimal passe effectivement par les μ-*agents* de plus forte fiabilité (c'est-à-dire A et D car 2*0.70 + 0.65 > 0.50 + 0.60 + 0.65).

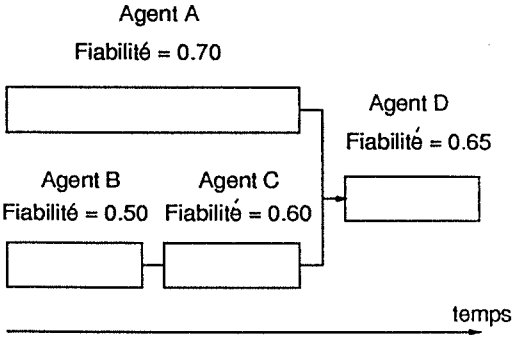


Figure 2: Recherche du chemin optimal

3.4 Le "moteur d'évolution"

Le chemin optimal (composé de phonèmes) est ensuite parcouru de façon à mettre, dans les zones de forte fiabilité, les phonèmes en relation avec les mots proposés par le système et à déterminer les zones de faible fiabilité dans lesquelles les Ω-*agents* doivent fournir de nouvelles solutions. Ceci est illustré par la figure 3. Dans les zones 1 et 2, la même

séquence de phonèmes est reconnue par les trois Ω-*agents* (avec tout de même un décalage temporel pour l'Ω-*agent* HMM_PH): on se trouve donc dans des zones de forte fiabilité dans lesquelles les mots reconnus sont: cent six mille. Dans la zone 3, les Ω-*agents* HMM_PH et GTP se sont mis d'accord pour reconnaître le mot zéro, dont on retrouve la décomposition phonétique quasi complète sur le chemin optimal. On peut donc en conclure que le mot zéro a été reconnu sur la zone 3. Ensuite, sur la zone 4, le mot zéro est de nouveau reconnu. Enfin, la zone 5 est une zone de faible fiabilité où les Ω-*agents* devront fournir de nouvelles solutions.

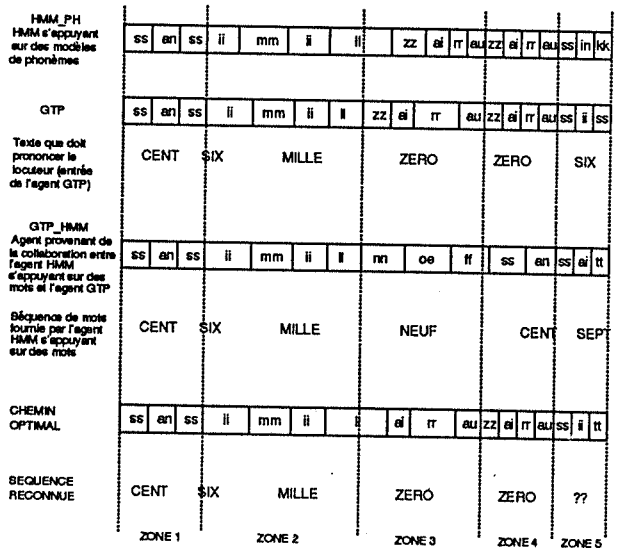


Figure 3: Recherche du texte prononcé par le locuteur

4 CONCLUSION

Il n'est pas actuellement possible de décrire les performances de $ETC_{vérif}$ car beaucoup de points doivent être améliorés (notamment le calcul de fiabilité et la coopération entre les phonèmes et les mots). Cependant, cette approche présente des caractéristiques intéressantes:

1. L'hétérogénéité des sources de connaissance – Actuellement, nous disposons de phonèmes et de mots. Par la suite, nous intégrererons également d'autres types d'informations comme des indices prosodiques [Lan95, LC95], des connaissances syntaxiques...

2. Un système distribué – Grâce au langage de programmation Oz, utilisé dans ce système, il est possible d'avoir des entités concurrentes (les μ-*agents*) avec des mécanismes de synchronisation.

3. L'interactivité – L'utilisateur peut devenir un Ω-*agent*: les solutions alors apportées sont prises en compte par le système (qui réagit à l'arrivée d'événements provoqués notamment par sélection de menus dans une interface graphique).

Cependant, des améliorations doivent encore être apportées sur ces différents points (comme, par exemple, la mise en relation des hypothèses) afin d'obtenir un système plus complet.

Bibliographie

[CF95] Jean-Luc Cochard and Philippe Froidevaux. Environnement multi-agents de reconnaissance automatique de la parole en continu. In *Actes des 3èmes Journées Francophones sur l'Intelligence Artificielle Distribuée et les Systèmes Multi-agents*, pages 101–110, mars 1995.

[CO95] Jean-Luc Cochard and Olivier Oppizzi. Reliability in a multi-agent spoken language recognition system. In *4th European Conference on Speech Communication and Technology*, Madrid, Spain, Sep 1995.

[HCA94] Béat Hirsbrunner, Michèle Courant, and Marc Aguilar. Parallelism and artificial intelligence. In *University of Fribourg Series in Computer Science Volume 3*, Institute of Informatics, University of Fribourg, Fribourg, Switzerland, August 1994.

[Lan94] Susan E. Lander. Distributed search and conflict management among reusable agents. Phd thesis, Dept of Computer Science, University of Massachusetts Amherst, May 1994.

[Lan95] Philippe Langlais. Traitement de la prosodie dans les systèmes de reconnaissance automatique de la parole. Thesis, Université d'Avignon et des pays du Vaucluse, France, October 11, 1995.

[LC95] Philippe Langlais and Jean-Luc Cochard. The use of prosodic agents in a cooperative automatic speech recognition system. In *International Congress of Phonetic Sciences*, Stockholm, Sweden, August 13–19 1995.

[Tor94] Kari Torkkola. New ways to use LVQ-codebooks together with hidden Markov models. In *ICASSP*, Adelaide, Australia, April 1994.

[Via95] Murielle Vial. Définition et évaluation d'un protocole de négociation dans un système multi-agents de reconnaissance de la parole. Technical report, IDIAP, Martigny, Switzerland, June, 1995.

GOBE-TOUT EN DÉTECTION DE MOTS NOUVEAUX ET EN DÉTECTION DE MOTS-CLÉS

Rachida EL MÉLIANI, Douglas O'SHAUGHNESSY

INRS-Télécommunications - 16 place du Commerce - Verdun (Ile des Soeurs) - H3E 1H6 CANADA

Tél.: (514) 761 8616 - Fax: (514) 761 8501 - e-mail: meliani@inrs-telecom.uquebec.ca

ABSTRACT

In this paper we propose different kinds of design for fillers when used for new-word detection as well as keyword spotting. The first type is a parallel model of all English phonemes the models of which are trained on out-of-vocabulary words. The second one is defined at the lexical level only: it uses the same HMMs of phonemes for vocabulary and non-vocabulary words; it consists of a parallel set of subunits (phonemes or syllables), which is used in the lexical graph and in the definition of language models. As for the last one, it uses combinations of both levels.

In spite of the obvious similarity between new-word detection and keyword spotting, our results show that for keyword spotting the set of lexical phonemic fillers performs very well, far better than the others, while in new-word detection the lexical syllabic filler is the one that gives the best performance.

1. INTRODUCTION

La limitation du vocabulaire est particulièrement contraignante lors de l'utilisation de systèmes de reconnaissance de la parole par le public et cela par comparaison aux nombreuses limitations imposées à ces systèmes par les limites de l'ordinateur. Des mots tronqués, des hésitations ou même des mots hors vocabulaire apparaissent même dans les discours des utilisateurs avertis de ces systèmes. Ceci définit un des problèmes-clés de la reconnaissance de la parole spontanée: «*comment distinguer de manière significative les mots du vocabulaire de ceux hors vocabulaire?*».

Selon l'application envisagée on distingue deux axes de traitement des mots hors vocabulaire:

- La détection des mots-clés [Rose, 1990; Rose, 1993] qui consiste en la détection des mots du vocabulaire prédéfini en rejetant les mots hors vocabulaire, quand on désire retrouver par thèmes l'information contenue dans le signal.

- La détection de mots nouveaux [Asadi, 1991; Asadi,1993] qui consiste en la détection de l'occurrence des mots hors vocabulaire. Elle s'accompagne généralement de la reconnaissance des mots du vocabulaire et propose éventuellement une transcription phonétique des mots hors vocabulaire. Elle est utilisée principalement dans les applications comme le dictaphone ou la reconnaissance à vocabulaire illimité.

Cet article présente différents nouveaux types de gobe-tout (filler) adaptés à la détection des mots-clés (que nous noterons désormais DMC) et à la détection de mots nouveaux (qui sera notée DMN), sur des systèmes utilisant le système de reconnaissance de la parole continue à grand vocabulaire de l'Institut National de la Recherche Scientifique en Télécommunications (INRS-Télécommunications) [Kenny,1994].

2. DESCRIPTION DU SYSTÈME

2.1. Le Système de Reconnaissance de l'INRS

Le système de DMC aussi bien que celui de DMN que nous avons construits sont basés sur le système de reconnaissance en temps réel de la parole continue à grand vocabulaire de l'INRS utilisant les HMMs [Kenny,1994]. L'avantage d'un tel choix est qu'il permet de faire varier la taille du vocabulaire prédéfini sans grande augmentation du temps de calcul donc de s'adapter à un plus grand choix d'applications.

Le schéma-bloc représentant ce système est donné en figure 1. La recherche de mot commence par une recherche arrière Viterbi sur le graphe phonétique dérivé de l'arbre lexical pour produire la table B* des estimées de tous les scores de segment de phonème ainsi que les fins de segment. Puis vient une recherche avant A* qui utilise la table B* pour produire

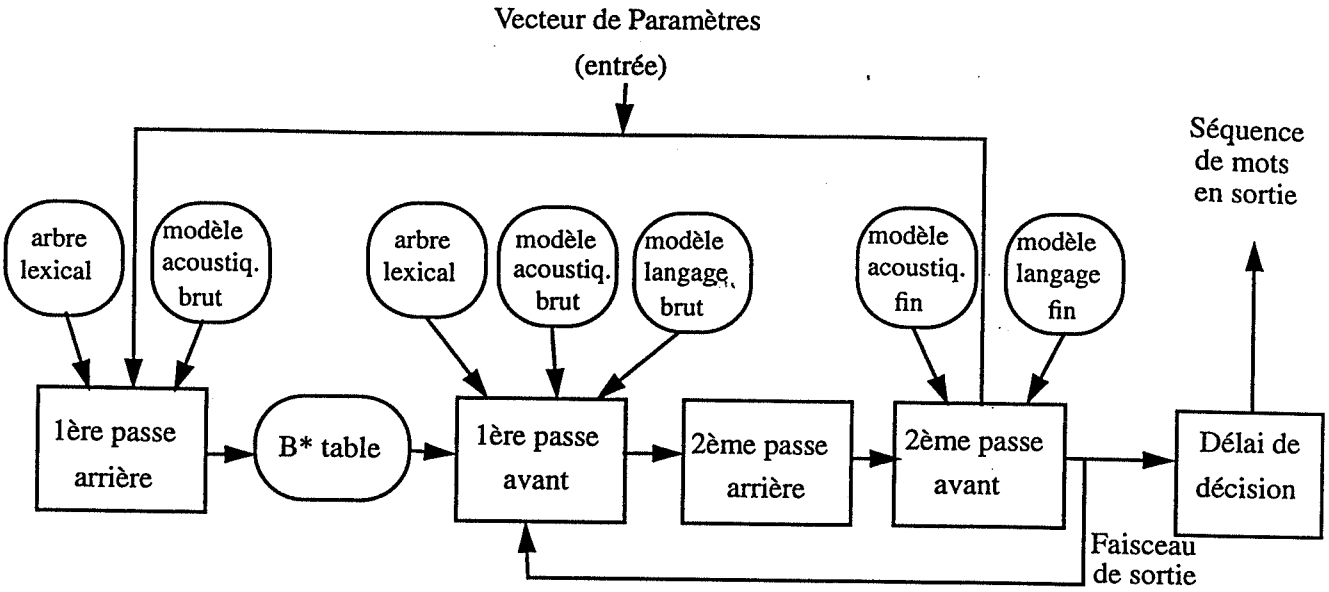


Figure 1: Le système de reconnaissance de l'INRS.

une probabilité surestimée des scores futurs pour réduire efficacement le rayon de recherche. Durant la seconde passe, le graphe restreint de mots est réévalué en utilisant des modèles HMMs de phonèmes dépendant du contexte et un modèle de langage raffinés.

2.2. Détection de Mots-clés versus Détection de Mots Nouveaux

À première vue, il existe une profonde similitude entre les deux problèmes, liée principalement à la dualité 'mots du vocabulaire/mots hors vocabulaire' qui suggère l'utilisation de gobe-tout dans les deux cas. Cependant il existe des différences importantes tant au sujet de la taille du vocabulaire, que de la nature des mots hors vocabulaire, ou de la précision de la détection.

3.2. Le Détecteur de Mots-clés

3.2.1 Gobe-tout acoustico-phonétiques

Dans un premier temps, on suppose que la représentativité des mots hors vocabulaire dans le corpus d'apprentissage est suffisante pour définir des HMMs de phonèmes spécifiques à ces mots. Ainsi ceux-ci seront modélisés à l'aide d'un gobe-tout constitué d'un classique réseau consistant en la mise en parallèle de tous les HMMs des phonèmes de l'anglais entraînés sur la parole hors vocabulaire (on le référence par gobe-tout acoustico-phonétique), alors que les HMMs des phonèmes de mots-clés sont entraînés sur toutes les occurrences de ces derniers. Cela nécessite, pour une détection adéquate des mots-clés, que les phonèmes dépendant du contexte les constituant soient chacun présent un grand nombre de fois dans l'ensemble des occurrences de mots-clés, d'où une dépendance à la fréquence et au nombre des mots-clés et donc une restriction au choix des mots-clés.

3.2.2. Gobe-tout Lexicaux

Pour se libérer de cette contrainte sur le corpus d'entraînement, un deuxième type de gobe-tout dit lexical a été utilisé: les HMMs de phonèmes sont cette fois-ci les mêmes pour les mots du vocabulaire et ceux hors vocabulaire, et la distinction entre ces deux types de mots se fait uniquement au niveau lexical (arbre lexical et modèle de langage). Les gobe-tout lexicaux utilisés pour représenter dans l'arbre lexical toute parole hors du vocabulaire consistent simplement en la mise en parallèle de sous-unités de mots (phonèmes ou syllabes).

3.2.3. Gobe-tout Mixtes

Un dernier type de gobe-tout mixtes combinant la définition des gobe-tout aux niveaux lexical et acoustico-phonétique a également été testé.

3.2.4. Modèle de langage

Le modèle de langage utilisé dans le système de reconnaissance de l'INRS assigne une probabilité à chaque chaîne de mots en tenant compte des fréquences des unigrammes et des

Nom	Type	Locu-teurs	Taille (mots)	Bruit
ATIS	Donnée de voyage	285 mixtes	1030	calme
WSJ	Lecture	1 mâle	4872	normal

Figure 2: Bases de données.

bigrammes des mots du vocabulaire. Ainsi en DMC ces fréquences sont calculées pour les gobe-tout sur les seules occurrences de mots hors du vocabulaire dans le corpus d'apprentissage alors que les fréquences de mots-clés sont calculées sur les occurrences de mots-clés dans le même corpus.

3.3. Le Détecteur de Mots Nouveaux

Le système de DMN que nous avons construit utilise des gobe-tout lexicaux d'architecture identique à ceux utilisés en DMC, cependant le modèle de langage utilisé calcule les fréquences des gobe-tout sur tous les mots du vocabulaire du corpus d'apprentissage puisqu'on ne possède pas d'information spécifique aux mots nouveaux.

3. TESTS ET RÉSULTATS

3.1. Les Vocabulaires Utilisés

Les résultats sont donnés pour des vocabulaires tirés de deux bases de données différentes (figure 2): Wall Street Journal (WSJ) et ATIS (Air Travel Information System). Les vocabulaires utilisés dans les expériences de DMC décrites ici sont tous tirés de Wall Street Journal, leurs tailles varient de 23 à 100 mots de fréquences variables.

Quant à la DMN, elle a été réalisée sur un vocabulaire de 1018 mots tirés de ATIS réservant ainsi 12 mots nouveaux, ainsi que sur un vocabulaire de 4842 mots de Wall Street Journal laissant ainsi 30 mots nouveaux puis sur un vocabulaire de 4654 mots connus de Wall Street Journal et donc 218 mots inconnus.

3.2. Tests en Détection de Mots-clés

3.2.1. Gobe-tout Acoustico-phonétique

Les résultats obtenus avec le gobe-tout phonétique varient beaucoup avec la fréquence des phonèmes dépendants du contexte constituant les mots-clés. Une fréquence de 30 au minimum est nécessaire pour une détection d'au moins 70%. Le choix des mots-clés est donc très contraint. Les résultats apparaissent en figure 3 où 'système total' représente le système de reconnaissance de base utilisé sur le vocabulaire complet (8472 mots) et simplifié pour pouvoir tourner sur une station DEC.

3.2.2. Gobe-tout Lexicaux

Les gobe-tout lexicaux permettent de se débarrasser de la contrainte précédente tout en conservant une précision convenable. La figure 3 reporte les résultats pour trois types de gobe-tout: un seul gobe-tout englobant tous les phonèmes de l'anglais (I), un ensemble de gobe-tout se partageant l'ensemble des syllabes du corpus (II), l'ensemble des phonèmes de l'anglais (un gobe-tout par phonème, noté III). Les résultats montrent la nette supériorité de ce der-

	Système total	Acoust. phonét.	Lexical I	Lexical II	Lexical III	Mixte I	Mixte II	Mixte III
Taux détection (%)	80	75	67,63	82,9	92	77	82,6	93
Fausses alarmes (fa/h/mc)	X	0,15	0,10	0,10	1,2	0,63	0,23	0,39

Figure 3: Détection de mots-clés

Vocabulaire	R (%)	DMN (%)	TMN (%)
WSJ	66	70	70

Figure 4: Détection de mots nouveaux pour le gobe-tout lexical phonémique unique.

nier ainsi que son efficacité.

3.2.3 Gobe-tout Mixtes

Les résultats de l'utilisation de ces gobe-tout lexicaux avec des modèles spécifiques aux mots hors du vocabulaire sont reportés en figure 3. Ces systèmes sont coûteux en mémoire et en temps mais apporte une substantielle amélioration dans la plupart des cas.

3.3. Tests en Détection de Mots Nouveaux

Les résultats sont donnés en figure 4. Les abréviations ont été utilisées pour alléger le tableau: R est mis pour taux de reconnaissance des mots du vocabulaire, DMN pour taux de détection des mots nouveaux et TMN pour taux de transcription des mots nouveaux.

3.3.1. Gobe-tout Lexical Phonémique unique

Les résultats obtenus pour le gobe-tout lexical Phonémique unique (figure 4) sont assez moyens du fait probablement de son manque de spécificité.

3.3.2. Gobe-tout Lexicaux Phonémiques

Même en séparant les phonèmes (figure 5), on n'obtient pas de grande amélioration à cause de la difficulté à leur associer dans le modèle de langage des fréquences adéquates.

3.3.2. Gobe-tout Lexicaux Syllabiques

Il donne les meilleurs résultats (figure 6). Cela doit être lié à la contrainte linguistique imposée par les syllabes.

Vocabulaire	R (%)	DMN (%)	TMN (%)
WSJ	65	68	64

Figure 5: Détection de mots nouveaux pour les gobe-tout lexicaux phonémiques

Vocabulaire	R (%)	DMN (%)	TMN (%)
WSJ	72	82	85
ATIS	78	80	65

Figure 6: Détection de mots nouveaux pour le gobe-tout lexical syllabique

4. CONCLUSION

Nous avons proposé différentes architectures de gobe-tout puis nous avons étudié leur effet en détection de mots-clés aussi bien qu'en détection de mots nouveaux. Malgré la similitude de ces deux types de traitement, les gobe-tout testés donnent des résultats différents pour chacun d'entre eux: ainsi en détection de mots-clés l'ensemble des gobe-tout lexicaux phonémiques se révèle être le plus efficace si l'on tient aussi compte de la consommation de temps (entraînement indépendant du vocabulaire) et mémoire alors qu'en détection de mots nouveaux l'ensemble des gobe-tout lexicaux syllabiques est celui qui donne les meilleurs résultats.

5. BIBLIOGRAPHIE

Kenny P., Boulianne G., Garudadri H., Trudelle S., Hollan R., Lennig M.and O'Shaughnessy D. (1994) "Experiments in Continuous Speech Recognition Using Books on Tape", *Speech Communication*, V. 14 n. 1, 49-60

Rose R.C. et Paul D.B. (1990) "A Hidden Markov Model Based Keyword Recognition System", *International Conference on Acoustic Speech and Signal Processing*, 129-132.

Rose R.C. et Hofstetter E.M. (1993) "Task Independent Wordspotting Using Decision Tree Based Allophone Clustering", *International Conference on Acoustic Speech and Signal Processing*, II-467-470.

Asadi A., Schwartz R.et Makhoul J. (1991) "Automatic Modeling for Adding New Words to a Large-vocabulary Continuous Speech Recognition System", *International Conference on Acoustic Speech and Signal Processing* , 305-308.

Asadi A.O. et Leung H.C. (1993) "New-Word Addition and Adaptation in a Stochastic Explicit-segment Speech Recognition System", *International Conference on Acoustic Speech and Signal Processing*, V-642-645.

INTRODUCTION DE PARAMETRES PHONETIQUES EN POST TRAITEMENT D'UN SYSTEME MARKOVIEN DE RECONNAISSANCE DE LA PAROLE

Katarina BARTKOVA, Denis JOUVET

France Télécom - CNET - Technopole Anticipa, 2, Av. Pierre Marzin - 22307 Lannion Cedex

Tél.: 96 05 10 58 - Fax: 96 05 35 30 - e-mail: bartkova@lannion.cnet.fr

ABSTRACT

This paper deal with the use of two phonetic parameters in a speech recognition process: the segmental duration and the degree of the sound voicing. These parameters are used in a post-processing procedure of a system based on a Markov modelling of the speech signal. The statistic modelling of the phonetic phenomena, using phonetic knowledge, provides a post-processing score which is used to rescore the N-best solutions given by the speech recognition system. The efficiency of these phonetic parameters is evaluated on speech corpora recorded through the telephone.

1. INTRODUCTION

La présente étude traite de l'utilisation de deux paramètres phonétiques en reconnaissance de la parole : la durée et le degré de voisement des segments phonétiques. Ces paramètres sont exploités en post-traitement des N meilleures solutions fournies par un système de reconnaissance de parole à base de modèles de Markov. Ces informations sont complémentaires des informations spectrales prises en considération par les modèles de Markov et leur modélisation permet de calculer un score de post-traitement qui est utilisé pour réordonner les N meilleures solutions. Ces informations supplémentaires servent donc à confirmer ou infirmer les solutions proposées par les modèles de Markov. La modélisation de ces informations est élaborée à partir de connaissances phonétiques et les paramètres des modèles sont statistiquement estimés sur des corpus d'apprentissage.

L'utilisation d'un modèle prédictif des durées segmentales à permis la prise en compte des durées segmentales dans le cadre d'un post-traitement des N meilleures solutions (Bartkova, 1995). Il existe également de nombreuses études concernant l'utilisation de la durée segmentale dans la modélisation markovienne. Quelques unes utilisent la durée minimale (Gupta, 1992) ou une durée segmentale dépendante de la vitesse d'articulation (Suaudeau, 1993). Cependant très peu d'études prennent en considération des connaissances phonétiques dans la modélisation même des durées phonémiques (Bartkova, 1995) (Dumouchel, 1995).

Le trait de voisement a été employé dès les années 50 dans les premiers systèmes de reconnaissance de la parole. Il était le plus souvent utilisé, conjointement à d'autres traits phonétiques, par des systèmes analytiques qui cherchaient à détecter les traits soit sur la trame acoustique même, soit sur les segments (Rabiner, 1976). Le trait de voisement a été également utilisé pour la vérification descendante afin de recalculer le score des détections lexicales (Vivès, 1981) ou pour la détermination des cohortes des mots les plus probables (Meloni, 1992).

2. MISE EN OEUVRE ET CORPUS

Le système de reconnaissance employé dans cette étude, Phil90, repose sur une modélisation markovienne du signal de parole. Il traite des trames acoustiques qui correspondent aux coefficients cepstraux et à leurs dérivées temporelles.

Pour les expériences décrites dans ce document, une modélisation allophonique des mots à reconnaître est employée, et l'algorithme de reconnaissance détermine pour chaque mot (ou phrase) les 5 meilleures solutions. La modélisation allophonique prend en considération les influences contextuelles des sons (Jouvet, 1991). L'alignement correspondant à chacune des 5 meilleures solutions définit les segments phonétiques de la solution, à partir desquels seront calculés les paramètres segmentaux.

Les scores de post-traitement sont calculés grâce à une double modélisation des informations liées aux segments (Lokbani, 1993). L'une correspond aux événements observés le

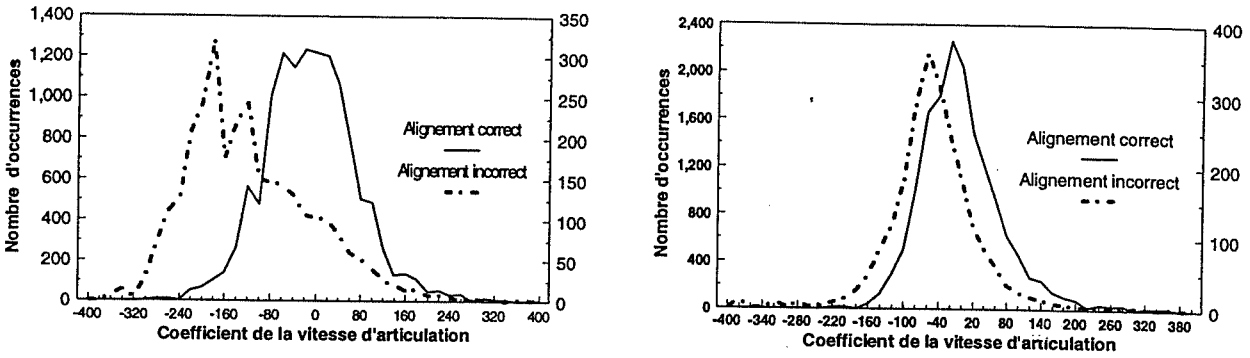


Figure 1: Histogrammes du coefficient de vitesse d'articulation des voyelles pour le corpus « Chiffres » (à gauche) et « Baladins2 » (à droite).

long des alignements corrects et l'autre aux événements observés le long des alignements incorrects.

Les évaluations sont menées sur divers corpus de parole téléphonique. Les corpus « Chiffres » et « Trégor », dont les vocabulaires correspondent respectivement aux 10 chiffres et à 36 mots et expressions, ont été enregistrés à travers le réseau téléphonique. 800 locuteurs ont prononcé tous les mots du vocabulaire. Le corpus « Baladins2 » correspond à de la parole spontanée. Il provient de 1000 appels à un serveur vocal acceptant des phrases en entrée et fonctionnant par détection de mots clé. Le vocabulaire de commande comprend 26 mots clé, dont une partie est présente dans le corpus Trégor. Ce mode de fonctionnement implique l'utilisation d'un modèle de rejet pour rejeter (et donc ignorer du point de vue du service) les mots non significatifs pour le dialogue. Le modèle de rejet employé dans ces expérimentations repose sur un ensemble de modèles acoustiques associés aux macro classes suivantes : voyelles orales, voyelles nasales, fricatives, occlusives, nasales, liquides et semi-voyelles.

Chaque corpus est découpé en deux parties ; l'une contribue à l'estimation des paramètres (des modèles de Markov - pour les corpus Chiffres et Trégor - et des modèles de post-traitement dans tous les cas) ; l'autre est réservée aux tests de reconnaissance (mesure des performances). En ce qui concerne les évaluations sur le corpus Baladins2, les modèles de Markov employés (pour la détermination des N meilleures solutions) ont été appris sur un autre corpus de parole correspondant à des prononciations isolés des mots clé du vocabulaire.

3. DUREE SEGMENTALE

Alors qu'une précédente étude évaluait l'apport de la durée segmentale (Bartkova, 1995) en post-traitement des N-meilleures solutions sur des mots isolés, nous l'appliquons ici sur de la parole continue. Chaque solution correspond donc à une suite de mots clé, et non plus à un seul mot. Une autre différence concerne l'emploi d'un modèle de rejet. Nous allons comparer les résultats obtenus sur la parole continue avec les résultats précédemment obtenus sur les corpus de mots isolés.

3.1 Prédiction de la durée phonémique

Le modèle de prédiction des durées employé prenait en considération des phénomènes phonétiques et phonologiques pertinents en français. Le modèle tenait compte des contextes gauches et des contextes droits pour les différents phonèmes, ainsi que de leur position dans la syllabe et dans le mot. A la différence du traitement des mots isolés, nous avons dû ajouter ici la distinction entre une dernière syllabe d'un mot suivie d'une pause et une dernière syllabe d'un mot suivie d'un autre mot. N'ayant pas à notre disposition la phrase complète (le système ne reconnaissait que les mots clé) il nous a été impossible de considérer la profondeur de la structure syntaxique exprimée par la durée de la dernière syllabe du mot. De ce fait on obtenait une plus grande variation de la durée sur la dernière syllabe que sur les autres syllabes.

Pour la parole spontanée, il nous a fallu modéliser également les durées des modèles de rejet (macro-classes de phonèmes). Le contexte n'a pas été pris en considération pour la prédiction des durées correspondant à ces unités (macro-classes).

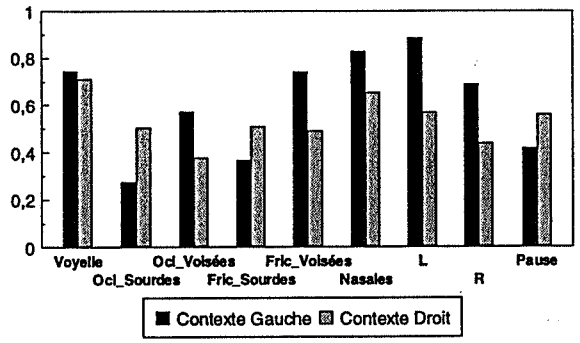


Figure 2 : Degré de voisement des voyelles dans différents contextes gauches et droits.

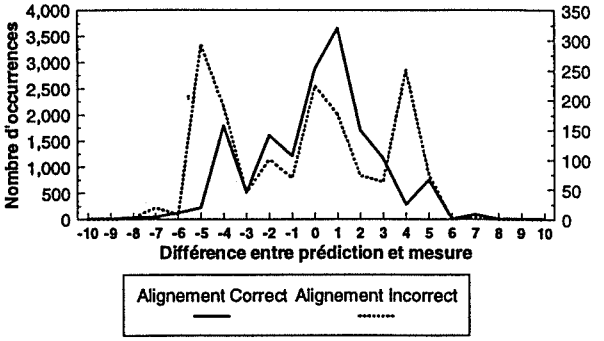


Figure 3 : Histogrammes des erreurs de prédiction du degré de voisement.

3.2 Modélisation de la vitesse d'articulation

Une des caractéristiques principales de la durée est l'élasticité: certains locuteurs parlent plus vite que d'autres. Aussi, pour tenir compte de ce phénomène, deux coefficients de vitesse d'articulation sont calculés pour chaque phrase : un pour les voyelles et un autre pour les consonnes. Un 3ème coefficient est calculé pour les modèles de rejet. Ces coefficients sont déterminés pour chaque phrase de manière à minimiser la somme des erreurs de prédiction des durées des segments de la phrase considérée.

Pour les mots isolés monosyllabiques, contenant une seule voyelle et une seule consonne l'erreur de prédiction devient nulle en raison de la détermination des coefficients de la vitesse d'articulation. Il est donc indispensable de tenir compte de ces coefficients de vitesse d'articulation dans le post-traitement. Pour cela une modélisation des ces coefficients est effectuée tant pour les alignements corrects que pour les alignements incorrects. La figure 1 qui représente les histogrammes correspondants aux alignements corrects et aux alignements incorrects montre que ceux-ci se différencient beaucoup plus nettement pour le corpus des mots isolés (à gauche) que pour le corpus de parole continue (à droite).

3.3 Evaluations

La table 1 indique le pourcentage de phrases (réduites à des mots isolés pour les corpus Chiffres et Trégor) pour lesquelles le score post-traitement des durées segmentales conduit à la bonne réponse (recherchée parmi les 5 meilleures solutions disponibles évidemment). L'utilisation de la durée (erreur de prédiction et vitesse d'articulation) donne des résultats comparables sur les différents corpus de parole.

La combinaison des scores liés aux durées avec les scores markoviens conduit à une réduction des taux d'erreur de 5 % sur le corpus des « Chiffres » et de 8 % sur le corpus « Trégor » par rapport à l'utilisation du modèle de Markov seul.

4. TRAIT DE VOISEMENT

Un degré de voisement a été modélisé pour les différents phonèmes qui constituent les mots clé. Il correspond au pourcentage de trames du segment phonétique qui ont été considérées comme voisées lors de la détermination du pitch.

4.1 Prédiction du degré de voisement

Pour prédire le degré de voisement des phonèmes constituant les mots clé, nous les avons regroupés en 13 macro-classes : 5 vocaliques et 8 consonantiques. Lors de la modélisation 13 contextes gauches et 13 contextes droits ont été pris en considération.

Les résultats illustrés par la figure 2 montrent que le contexte gauche a une influence

Table 1 : Pourcentage de phrases pour lesquelles le score de post-traitement des durées segmentales conduit à la bonne réponse (recherchée parmi les 5 meilleures solutions disponibles).

	Chiffres	Trégor	Baladins2
Erreur de prédiction	45 %	64 %	60 %
Vitesse d'articulation	68 %	51 %	67 %
Erreur de prédiction + Vitesse d'articulation	76 %	80 %	75 %

phonétiquement cohérente sur le degré de voisement du phonème suivant : un phonème non-voisé abaisse le degré de voisement du phonème suivant et un phonème voisé l'augmente (assimilation proactive du trait de voisement). L'influence du contexte droit est moins claire.

4.2 Utilisation dans le post-traitement

Le paramètre modélisé dans le post-traitement est l'erreur de prédiction du voisement : c'est-à-dire l'écart entre le degré de voisement prédit par le modèle et le degré de voisement mesuré sur le segment.

La figure 3, qui représente les histogrammes des erreurs de prédiction du degré de voisement des plosives voisées, illustre la capacité de discrimination de ce paramètre entre les alignements corrects et les alignement incorrects.

4.3 Evaluations

Les premiers résultats, obtenus en utilisant le degré de voisement dans le post-traitement sont très encourageants. Le tableau II indique l'évolution du taux de bonne reconnaissance pour différentes combinaisons du score de post-traitement associé au degré de voisement avec le score markovien (la 1ère ligne indique la contribution du score de voisement dans le calcul du score final). La solution (parmi les 5 meilleures) qui obtient le meilleur score final détermine alors la réponse du système global. Dans l'ensemble, le trait de voisement permet une réduction de 21% du taux d'erreur de reconnaissance des phrases. Mais il est utile de mentionner que, pour ce corpus, il y a plus de 20 % d'erreurs irrécupérables par le post-traitement (la réponse correcte ne se trouvant pas parmi les 5 meilleures solutions).

5. CONCLUSION

Nous avons présenté l'utilisation de deux paramètres phonétiques en post-traitement d'un système de reconnaissance automatique de la parole basé sur les modèles de Markov. Le rôle de ces paramètres phonétiques est de confirmer ou de pénaliser les solutions proposées par le décodage markovien. La prise en compte de ces informations dans le post-traitement conduit à réordonner les N-meilleures solutions et permet de récupérer certaines erreurs commises par le système de reconnaissance purement markovien.

Les résultats obtenus ici montrent que l'utilisation des connaissances phonétiques dans les systèmes automatiques de reconnaissance de la parole permet une amélioration des performances.

6. REFERENCES

Bartkova K., Jouvet D., Moudenc T. (1995), Using Segmental Duration Prediction for Rescoring the N-Best Solutions in Speech Recognition, *XIIIth ICPhS*, août 1995 Stockholm, Suède, Vol. 4, pp. 248-251.

Gupta V. et al. (1992), Use of minimum duration and energy contour for phonemes to improve large vocabulary isolated-word recognition, *Computer Speech and Language*, Vol. 6, pp. 345-359.

Suaudeau N. & André-Obrecht R. (1993), Sound duration modelling and time-variable speaking rate in a speech recognition system, *EUROSPEECH*, septembre 1993, Berlin, Allemagne, pp. 307-310.

Dumouchel P. & O'Shaughnessy D. (1995), Segmental Duration and HMM Modelling, *EUROSPEECH*, septembre 1995, Madrid, Espagne, pp. 803-805.

Rabiner L.R. & Sambur M.R. (1976), Some Preliminary Results on the Recognition of Connected Digits, *IEEE Trans. on Acoustics, Speech and Signal Processing*, Vol. 24, n° 2.

Vivès R. (1981), Vérification des hypothèses proposées par l'analyseur lexical d'un système de reconnaissance automatique de la parole, *3th FASE symposium*, Avril, Venise, Italie, pp. 277-280.

Meloni H., Bechet F. & Gilles P. (1992), Reconnaissance analytique de mots isolés d'un grand lexique, *19èmes JEP*, mai, Bruxelles, Belgique, pp. 195-199.

Jouvet D., Bartkova K. & Monné J. (1991), On the modelisation of allophones in an HMM based speech recognition system, *EUROSPEECH*, septembre 1991, Gènes, Italie, pp. 923-926.

Lokbani M.N., Jouvet D. & Monné J. (1993), Segmental post-processing of the N-best solutions in a speech recognition system, *EUROSPEECH*, septembre 1993, Berlin, Allemagne, pp. 811-814.

Table II : Taux de reconnaissance sur le corpus Baladins2 après prise en compte du degré de voisement dans le post-traitement des 5 meilleures solutions.

	Voisement seul	Modèles de Markov et voisement				Markov seul
Contribution voisement	(i.e. 100 %)	15 %	10 %	5 %	3 %	(i.e. 0 %)
Ensemble apprentissage	62,1 %	65,5 %	66,4 %	67,8 %	**68,2 %**	57,9 %
Ensemble de test	59,1 %	63,4 %	64,8 %	66,7 %	**67,9 %**	59,3 %

MISE EN ŒUVRE DES RÉSEAUX DE NEURONES GAMMA POUR LA SEGMENTATION DE LA PAROLE CONTINUE

Laurent Buniet buniet@loria.fr
Dominique Fohr fohr@loria.fr
Jean-Marie Pierrel jmp@loria.fr

CRIN-CNRS & INRIA Lorraine - Campus scientifique Victor Grignard - B.P. 239 - F-54506 Vandœuvre-lès-Nancy

Tél : 83 59 20 31 - Fax : 83 41 30 79

ABSTRACT

This paper presents the application of a locally recurrent globally feedforward neural network model, the Gamma Memory Model, to the problem of speech segmentation. The Gamma Memory Model is composed of a local and constrained feedback. This feedback implements a infinite impulse response (IIR) filter which can model delayed inhibitions or excitations by exponential decay. The network can, thus, make a decision according to the current input and also by taking into account previous events which had a certain duration. Therefore, this architecture seems to be well suited to the problem of speech segmentation.

In this paper, we first present the tests made with gamma units in the input plane and standard neurons in the hidden and output layers. Then, we present the tests made with gamma neurons in the hidden layers : the architecture of the neurons in the hidden layer is modified, a gamma neuron being composed of a standard neuron followed by a gamma unit.

1. SEGMENTATION DE LA PAROLE CONTINUE

Lors de travaux antérieurs [2], nous avons essayé de mettre en place une méthode pour la reconnaissance de petits vocabulaires prononcés de manière continue en milieu bruité. Cette méthode reposait sur un enchaînement de trois perceptrons multicouches, chacun étant dédié à une tâche. Le premier était chargé de trouver les parties voisées du signal, ces parties servant de points d'ancrage au deuxième perceptron chargé de reconnaître les voyelles prononcées. Le troisième perceptron utilisé dans la chaîne des traitements permettait lui de reconnaître les mots dans le cas où la voyelle ne permettait pas seule de reconnaître ce mot (cas de «1» et «5» en français, par exemple).

Cette méthode avait de très bons résultats, même à de faibles rapports signal sur bruit (RSB), mais devenait inutilisable à des RSB nuls ou négatifs car le niveau de segmentation n'était plus capable d'effectuer son travail. Les sorties du réseau dédié devenaient entâchées d'erreurs et les points d'ancrage avaient des durées très peu plausibles [3].

Il aurait été possible de se baser sur des durées moyennes [7] mais cette connaissance a priori n'aurait pas amélioré la qualité du réseau et aurait donc été difficilement applicable.

Pour résoudre ce problème, nous avons donc essayé d'implanter un modèle qui soit capable de prendre en compte une notion de durée. Des recherches ont déjà été menées avec succès sur des architectures fortement récurrentes [11] mais nous avons choisis de nous orienter vers un modèle plus exploré mathématiquement et capable de simuler des inhibitions ou des excitations retardées. Nous nous sommes donc orientés vers un modèle possédant une décroissance exponentielle : le modèle gamma.

2. LES RÉSEAUX DE NEURONES GAMMA

Le modèle gamma a tout d'abord été présenté dans [6]. Il s'agit d'un filtre à réponse impulsionnelle infinie dont le comportement est fonction d'un coefficient μ responsable d'une conservation plus ou moins grande du signal. L'équation de ce filtre est donné par :

$$y_{i,t} = (1-\mu)\, y_{i,t-1} + \mu y_{i-1,t}$$

où $y_{i,t}$ représente la sortie de l'unité de rang i au temps t. Le filtre gamma a été initialement utilisé pour réaliser le «Focused Gamma Network» [9]. Dans cette architecture, les unités gamma ne sont présentes que dans la couche d'entrée et possèdent toutes le même coefficient μ. Le modèle gamma fait partie de la classe plus générale des perceptrons à récurrence locale [12].

Nous avons étendu l'architecture de deux manières (figure 1) [4]. Premièrement, nous

avons relâché les contraintes sur μ en autorisant à ce coefficient d'être soit spécifique à toutes les unités d'une même ligne de délais encastrés, soit d'être spécifique à chacune des unités gamma de la couche d'entrée, cette augmentation du nombre de degrés de liberté devant permettre au réseau de mieux approximer le problème.

Deuxièmement, nous avons modifié la sortie des neurones des couches cachées en ajoutant un filtre gamma à la sortie de chacun de ces neurones. Nous permettons ainsi au réseau de conserver une trace du signal antérieur, dans la couche d'entrée, mais aussi d'avoir une trace de ses activations antérieures, dans les couches cachées. La mémoire des couches cachées est

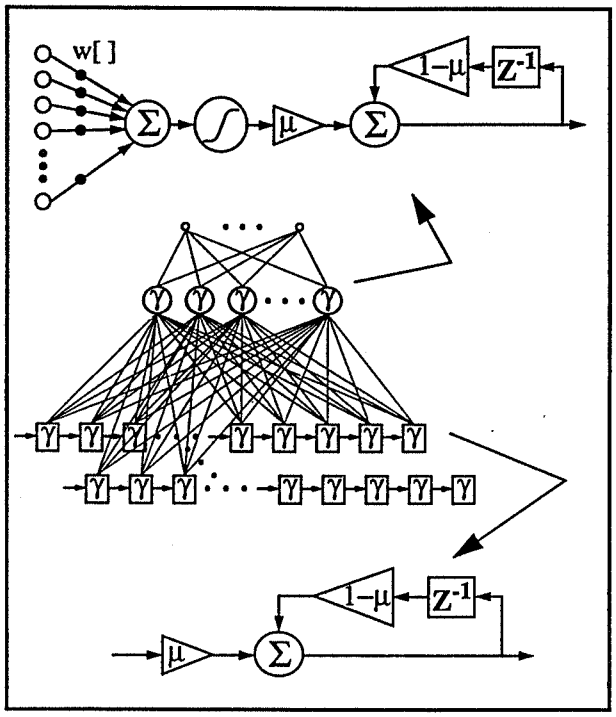


Figure 1: Un réseau de neurones gamma

donc plus abstraite et représente une sorte de mémoire à court terme des décisions du réseau.

Il est mathématiquement prouvé que le filtre gamma est stable pour des valeurs de μ comprises entre 0 et 2. Nous avons restreint cet intervalle à [0,1] de manière à n'avoir qu'un filtre passe-bas. Cette restriction des valeurs de μ nous permet d'avoir une interprétation directe des capacités de mémorisation de chaque neurone même si le rôle effectif du neurone reste sans explication.

La mise à jour des poids du réseau et la mise à jour des coefficients μ des neurones sont effectuées conjointement grâce à une procédure d'apprentissage. Cette procédure a été réécrite à partir de la rétropropagation du gradient d'erreur tel que cela avait été montré par Almeida [1] et Pineda [8]. Cette méthode n'est pas la meilleure puisque l'erreur est calculée à partir d'un gradient approximé [14] mais elle a à son avantage le fait de converger assez facilement, le modèle gamma étant parfois dur à mettre en œuvre [10]. Nous testons actuellement un algorithme d'apprentissage basé sur la rétropropagation dans le temps (BPTT, [13]).

La difficulté de l'apprentissage vient du fait que chaque neurone du réseau conserve des composantes du signal en cours de traitement, ces composantes étant difficiles à prendre en compte lors de l'apprentissage de la tâche. Ces composantes sont appelées les moments de Poisson du signal [5].

3. Résultats expérimentaux

3.1. Corpus utilisés

Le corpus utilisé lors des expérimentations est la base de données TIMIT. Nous avons sélectionné le partie «SI, diverse sentence type». Le corpus d'apprentissage est constitué de 24 phrases prononcées chacune par un locuteur masculin différent et extraites des répertoires «train/dr1» et «train/dr2». Le corpus de test est lui constitué de 75 phrases extraites des répertoires «test/dr1» et «test/dr2», plusieurs phrases pouvant être prononcées par un même locuteur masculin.

Sur les phrases sélectionnées, nous avons calculé des vecteurs de 12 coefficients MFCC, l'intervalle entre 2 vecteurs consécutifs étant de 4 millisecondes (ms). Cet intervalle de 4 ms, qui est faible relativement aux 10 ms d'intervalle que l'on trouve généralement dans la littérature, a été choisi pour amoindrir la fenêtre de signal observable par le réseau de neurones et donc l'obliger à implanter de la mémoire par le biais du filtre gamma.

L'étiquetage manuel des phrases a servi à préparer deux types d'expériences. La première expérience a consisté à reconnaître les voyelles dans le signal. La seconde expérience a consisté à reconnaître les parties vocaliques du signal, étant classées vocaliques les voyelles et les semi-consonnes.

3.2. Architecture du réseau

L'architecture du réseau utilisé pour les

expériences présentées est la même tout au long des différentes expériences. La plaque d'entrée du réseau est constituée de 12 lignes de 6 délais encastrés, à la manière d'un TDNN. L'unique couche cachée est constituée de 9 neurones. Au cours de nos différentes expériences, nous avons seulement fait varier la contrainte de partage des valeurs de μ ce qui a conduit aux conditions suivantes :

- la plaque d'entrée possédait soit 1 coefficient, soit 12 (1 coefficient par ligne de délais), soit 72 (1 par unité de la plaque),
- les neurones de la couche cachée étaient soit des neurones gamma, soit des neurones standards (la valeur de μ étant fixée à 1 tout au long de l'expérience, le comportement du neurone est alors identique à celui du neurone de McCulloch et Pitts).

3.3. Obtention des résultats

Les résultats sont donnés à deux niveaux : le niveau vectoriel et le niveau segmental.

Au niveau vectoriel, la réponse du réseau est considérée exacte si elle correspond à la sortie attendue.

Pour le calcul des résultats au niveau segmental, la sortie du réseau est d'abord lissée par une moyenne mobile sur 10 ms puis les résultats sont comparés à l'étiquetage manuel. Cinq pourcentages sont alors obtenus : les segments corrects, les segments insérés, les segments élidés, les segments fusionnés (lorsque plusieurs segments de l'étiquetage manuel sont regroupés dans un même segment par le réseau) et les segments divisés (lorsqu'un segments de l'étiquetage manuel est découpé en plusieurs segments par le réseau). Ces résultats étant calculés par la mise en correspondance de l'étiquetage manuel et de l'étiquetage obtenu avec le réseau de neurones, la somme des pourcentages dépasse 100%.

3.4. Résultats au niveau vectoriel

Les résultats au niveau vectoriel concernant les expériences de segmentation des voyelles ou de segmentation des noyaux vocaliques sont donnés dans les tables 1 et 2 respectivement. Les deux premières colonnes des tables 1 et 2 donnent des indications sur le nombre de degrés de liberté laissés au réseau, tant dans la plaque d'entrée que dans la couche cachée, la plaque d'entrée étant constituée de 12 lignes de 6 délais encastrés et la couche cachée de 9 neurones gamma. Les résultats ne présentent pas de différences significatives et doivent être étudiés au niveau segmental.

Table 1 : Résultat au niveau vectoriel pour la classification de type vocalique

Nombre de facteurs, couche d'entrée	Nombre de facteurs, couche cachée	Apprent., correct	Test correct
1	9	85 %	83 %
12	9	88 %	86 %
72	9	86 %	85 %

Table 2 : Résultat au niveau vectoriel pour la classification de type voyelle

Nombre de facteurs, couche d'entrée	Nombre de facteurs, couche cachée	Numéro de la tâche	Apprent., correct	Test, correct
12	0	#1	84 %	84 %
72	0	#2	85 %	84 %
72	9	#3	84 %	84 %

Un des faits intéressants à observer est la manière dont l'algorithme d'apprentissage a défini les valeurs des différents coefficients d'autorégression au sein de la plaque d'entrée. La figure 2 présente un graphique des valeurs

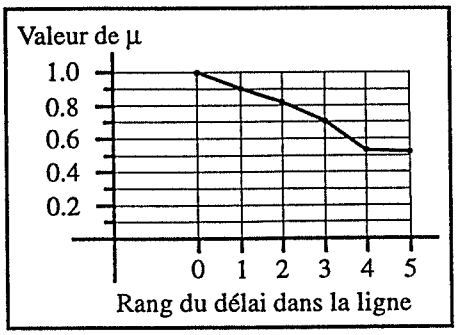


Figure 2: La valeur moyenne de μ comparée au rang du délai dans la ligne

moyennes de ces coefficients pour chacun des éléments de la ligne de délais. L'algorithme a adapté les délais de la ligne de manière à ce que ceux-ci mémorisent de plus en plus le signal à mesure que celui-ci parcourt la ligne de délais. Ainsi le signal de sortie de la première unité gamma de la plaque d'entrée est pratiquement le signal d'entrée alors que la dernière unité gamma de la ligne stocke à partie égale le signal du dernier pas de temps et le signal du pas de temps précédent. Ce stockage du signal est progressif et permet donc au réseau d'avoir

tout aussi bien une vue précise sur le signal d'entrée qu'une vue sur un signal composite constitué des traces des signaux d'entrée antérieurs.

3.5. Résultats au niveau segmental

La table 3 donne les résultats au niveau segmental. Les résultats montrent que le réseau

Table 3 : Résultat au niveau segmental pour la classification de type voyelle

	Test #1 12-0	Test #2 72-0	Test #3 72-9
correct	81 %	84 %	78 %
insertion	3 %	4 %	5 %
élision	10 %	12 %	18 %
fusion	9 %	5 %	4 %
division	3 %	3 %	1 %

peut tirer partie d'un grand nombre de degrés de liberté puisque les résultats s'améliorent lorsque l'on passe de l'architecture 1 à l'architecture 2.

Une augmentation supplémentaire du nombre de degrés de liberté n'a cependant pas permis d'améliorer les résultats. Le passage de l'architecture 2 à l'architecture 3 a fait chuter les résultats puisque le pourcentage de réponses correctes chute alors que le pourcentage des élisions augmente. Ce fait laisse à penser que l'algorithme d'apprentissage tel qu'il est dérivé de la rétropropagation simple n'est pas optimal. Il faut cependant noter que les pourcentages des fusions et des divisions décroissent tout au long des trois expériences.

4. Conclusion

Le modèle gamma permet de mémoriser des traces des signaux d'entrée grâce à un mécanisme de décroissance exponentielle qui peut être adapté à la tâche à approximer. Ce mécanisme devrait permettre de correctement modéliser les problèmes de segmentation mais également les problèmes de classification plus généraux. L'algorithme d'apprentissage pose encore un problème car une unité gamma est très fortement influencée par la trace du signal qu'elle traite. L'algorithme mis en œuvre ici n'est pas encore optimal. L'algorithme de la BPTT tenant compte de la nouvelle architecture des neurones a été implanté mais non encore suffisament testé.

5. Bibliographie

[1] Almeida L. B., «A learning rule for asynchronous perceptrons with feedback in a combinatorial environment», IEEE ICNN 87, pp 609-618, 1987

[2] Buniet L., Fohr D., Anglade Y., Junqua J.-C., Pierrel J.-M., «Selectively Trained Neural Networks for Connected Word Recognition in Noisy Environments», Eurospeech 93, pp 841-844, sep. 1993

[3] Buniet L., Fohr D., Junqua J.-C., «Une méthode connexionniste pour la reconnaissance de mots enchaînés en milieu bruité», Conférence Neurosciences et Sciences de l'Ingénieur, pp 27-30, mai 1994

[4] Buniet L., Fohr D., «Continuous Speech Segmentation with the Gamma Memory Model», Eurospeech 95, pp 1685-1688, sep. 1995

[5] Celebi S., «Representation of locally stationary signals using lowpass moments», PhD thesis de l'Université de Floride, sep. 1995

[6] De Vries B., Principe J. C., «The Gamma Model - A New Neural Model for Temporal Processing», Neural Networks, vol. 5, pp 565-576, 1992

[7] Junqua J.-C., «A Duration Study of Speech Vowels Produced in Noise», ICSLP 94, pp 419-422 juil. 1994

[8] Pineda F. J., «Generalization of Back-Propagation to Recurrent Neural Networks», Physical Review Letters, vol. 59, n° 19, pp 2229-2232, nov. 1987

[9] Principe J. C., De Vries B., Kuo J.-M., De Oliveira P. G., «Modeling applications with the focused gamma net», NIPS 4, pp 143-150, 1993

[10] Renals S., Hochberg M., «Using Gamma Filters to Model Temporal Dependencies in Speech», ICSLP 94, pp 1491-1494 juil. 1994

[11] Robinson A. J., «An application of recurrent nets to phone probability estimation», IEEE Trans. on Neural Networks, vol. 5, n° 2, pp 298-305, mar. 1994

[12] Tsoi A. C., Back A. D., «Locally Recurrent Globally Feedforward Networks: A Critical Review of Architectures», IEEE Trans. on Neural Networks, vol. 5, n° 2, pp 229-239, mar. 1994

[13] WerbosP. J., «Backpropagation Through Time: What it does and how to do it», Proceedings of the IEEE, vol. 78, n° 10, pp 1550-1560, oct. 1990

[14] Williams R. J., Zipser D., «A Learning Algorithm for Continually Running Fully Recurrent Neural Networks», Neural Computation, vol. 1, pp 270-280, 1989

Pour un système hybride de reconnaissance automatique de la parole continue

Stéphane Igounet

Laboratoire Informatique d'Avignon - 339, Chemin des Meinajariès - BP 1228 - 84911 Avignon Cedex 9 - France

Tél : (33) 90 84 35 35 - Fax : (33) 90 84 35 01 - e-mail : igounet@univ-avignon.fr

ABSTRACT

In the field of automatic continuous speech recognition, we describe a recognition system prototype. Then we propose a merging process for heterogeneous technique such as hidden Markov model, neural network and set of rules in synchronous and hybrid systems. Derived from the classical approach, the merging process can be directly integrated in decoding algorithm. In this view, we suggest a modification of the token passing model strategy.

KEYS WORDS

Automatic speech recognition, continuous speech, hidden Markov model, hybrid system, langage model.

RESUME

Dans le cadre de la reconnaissance automatique de la parole continue, nous décrivons le prototype d'un système de reconnaissance. Puis, nous proposons un moyen de fusionner des techniques hétérogènes telles que des chaînes de Markov cachées, des réseaux de neurones et des ensembles de règles dans des systèmes hybrides et synchrones. Dérivée de l'approche statistique traditionnelle, la fusion peut alors être directement intégrée dans l'algorithme de décodage. Dans cette optique, nous proposons une modification de la stratégie du *Modèle de Propagation de Jeton.*

MOTS CLES

Reconnaissance automatique de la parole, parole continue, modèles de Markov cachés, système hybride, modèles de langage.

1. INTRODUCTION

La réalisation d'un système de reconnaissance automatique de la parole continue nécessite de disposer d'un vaste ensemble de compétences. Du traitement du signal aux niveaux linguistiques les plus haut, les systèmes actuels intègrent de plus en plus de connaissances hétérogènes et ont demandé de nombreuses années de recherche et de développement. La plupart de ces systèmes sont plus ou moins fondés sur les mêmes principes [1, 2]. L'accent étant mis sur leur évaluation, l'objectif de réduction du taux d'erreur de reconnaissance est devenu le censeur de toute innovation majeure [1]. Notre travail s'inscrit dans cette problématique. Nous tentons de mettre au point un système qui prendrait en compte "l'état de l'art", tout en proposant des alternatives originales.

Dans la section 2, nous décrivons le prototype d'un système classique de reconnaissance automatique de la parole continue en différenciant les outils de mise au point du système réalisé. Puis, dans la section 3, nous présentons une modification possible de la stratégie de décodage afin de permettre la fusion de techniques hétérogènes dans un processus synchrone. Enfin, dans la conclusion, nous dégageons certaines pistes prometteuses.

2. PROTOTYPE DU SYSTEME DE RECONNAISSANCE

Notre objectif est de construire un système de reconnaissance automatique de la parole de type machine à dictée qui devra, à terme, répondre aux caractéristiques suivantes :

- Indépendance vis à vis de la langue (le français est privilégié).
- Parole continue.
- Multi-locuteur.
- Grand vocabulaire.

Dans un premier temps, nous décrivons les outils réalisés afin d'automatiser les différentes étapes de la mise au point. Puis, le prototype du système est présenté dans sa version actuelle.

Soulignons que les procédures de paramétrisation, d'apprentissage et de décodage qui nous

ont permis de réaliser ce système dérivent peu ou prou de l'ensemble logiciel HTK V1.5 [3].

2.1. Outils de mise au point

Ces modules nous permettent d'intégrer facilement et automatiquement la plupart des modifications faites sur le système. Ainsi, par exemple, la modification d'une règle phonologique dans le module de phonétisation est prise en compte dans les corpus et le lexique, mais aussi, lors de l'alignement automatique et donc de l'apprentissage.

- **Module de phonétisation**
  Nous utilisons le système GRIPHON [4] qui effectue une transcription automatique graphème-phonème pour phonétiser les phrases des corpus et les mots du lexique. Ce système est notamment fondé sur une base de règles (environ un millier de règles) et un module d'étiquetage grammatical statistique.
- **Module base de données**
  Gestion d'une base de donnée de parole (des phrases issues du journal Le Monde) constituée du signal, de la forme orthographique et de la forme phonétique.
  Nous disposons actuellement d'un peu plus de 1000 phrases en français de parole lue (pour le moment par un seul locuteur), réparties équitablement dans les corpus d'apprentissage (514 phrases, 19501 phonèmes) et de test (500 phrases, 21487 phonèmes).
- **Module d'alignement automatique multi-locuteurs**
  A partir du signal et des transcriptions phonétiques associées, nous pouvons réaliser un alignement du signal sur les phonèmes (un mécanisme de gestion des variantes phonologiques est opérationnel mais, il n'est pas encore véritablement utilisé dans la mesure où le module de phonétisation ne propose pas encore ces variantes). C'est cet alignement qui nous permet d'apprendre nos modèles acoustiques[1].

2.2. Description du système de RAP

Les principales caractéristiques du système actuel sont les suivantes :

- **Traitement du signal**
  Le signal est échantillonné à 16 Khz et caractérisé par un ensemble de paramètres calculés sur une fenêtre de Hamming d'une durée de 25 ms. Nous avons opté pour des vecteurs centisecondes composés de 12 coefficients cepstraux répartis sur une échelle Mel (MFCC) auxquels est ajoutée l'énergie. Le calcul de la dérivée et de l'accélération de ces 13 variables porte à 39 le nombre de composantes de chaque vecteur.
- **Modèles acoustiques**
  Nous disposons de deux types de modèles acoustiques : des modèles phonétiques et des modèles de diphones classifiés en contexte-droit [5]. Ils sont modélisés par des sources de Markov (de type modèle de Bakis) ayant 3 états émetteurs. Les probabilités d'émission des observations sont obtenues par des mixtures de gaussiennes dont les matrices de covariance sont supposées diagonales.
- **Apprentissage des modèles**
  L'apprentissage des modèles se décompose en trois phases. Une phase d'initialisation et deux phases distinctes d'estimation des probabilités avec l'algorithme de Baum-Welch. Lors de la phase d'initialisation, on procède au calcul des moyennes et des variances initiales en fonction des observations par le biais d'un processus itératif d'alignement et de calcul. La seconde phase consiste à appliquer localement l'algorithme de Baum-Welch en prenant en compte les frontières des unités. Enfin, la troisième phase applique de nouveau l'algorithme de Baum-Welch mais globalement, c'est-à-dire sans tenir compte de la segmentation (en mettant en cause les frontières des mots).
- **Décodage et Reconnaissance**
  Le décodeur utilise l'algorithme de Viterbi à travers la stratégie dite *Modèle de Propagation de Jeton* [6, 2]. L'espace de recherche, défini par l'intermédiaire d'une grammaire, est parcouru de manière synchrone et un modèle de langage bigramme peut être ajouté.

2.3. Premiers résultats

L'évaluation de notre système pose de nombreux problèmes. Tout d'abord, nous sommes

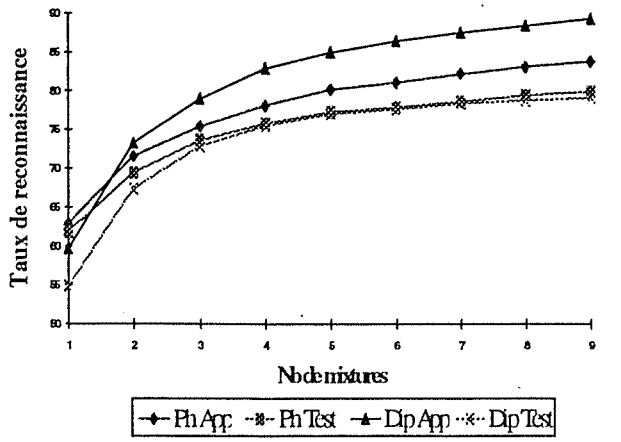
[1] Ce module d'alignement a été réalisé à partir d'un DAP multi-locuteur utilisant un modèle entraîné sur des mots isolés.

en phase de développement du système de base et nous devons encore y ajouter d'autres possibilités afin de pouvoir véritablement l'évaluer. Mais, le soin apporté à l'élaboration des différents modules nous permet de penser que ces améliorations pourront être mises en œuvre rapidement. De plus, nous sommes confrontés au problème de l'inexistence de corpus significatif en français et la solution adoptée (nous enregistrons nos propres corpus) empêche de véritables comparaisons.

Nous donnons, dans le tableau 1, les premiers résultats[2] obtenus pour un locuteur avec des modèles phonème et diphone sur un corpus d'apprentissage de 514 phrases et un corpus de test de 500 phrases. Nos tests ont été réalisés sans aucun modèle de langage.

Tableaux 1 : Premiers résultats.

	1	2	3	4	5	6	7	8	9
Ph App	62,90	71,52	75,41	78,11	80,20	81,07	82,21	83,12	83,83
Ph Test	61,82	69,42	73,65	75,85	77,31	77,86	78,62	79,43	79,93
Dip App	59,63	73,29	79,01	82,95	85,00	86,46	87,52	88,43	89,34
Dip Test	54,87	67,38	72,86	75,50	77,11	77,66	78,40	78,79	79,13



Comme on pouvait s'y attendre, nous obtenons une augmentation des taux de reconnaissance avec les diphones classifiés par rapport aux phonèmes. Mais de nombreux modèles acoustiques sont sous-représentés ce qui pose de nombreux problèmes lors de l'apprentissage et pour la comparaison des résultats[3].

[2] Le taux de reconnaissance prend en compte les insertions, les substitutions et les suppressions.

[3] Le passage de 35 machines phonétiques à 165 machines de diphones classifiés rend la sous-représentation des modèles encore plus importante.

3. FUSION DE CONNAISSANCES

La réalisation de systèmes de reconnaissance automatique de la parole demande de combiner différentes sources et niveaux de connaissances (acoustique, linguistique, phonétique, phonologique, sémantique, etc.) au moyen de stratégies de décodage *ad hoc*. Deux voies s'offrent à nous. Ou bien utiliser simultanément l'ensemble des connaissances disponibles et il s'agit alors de stratégies dites synchrones. Ou bien, on les exploite en couches successives de façon asynchrone [7, 2]. De plus, nous avons le choix entre plusieurs techniques (HMM, RNM, DTW, etc.) et nous souhaiterions pouvoir les fusionner dans des systèmes hybrides.

3.1. Présentation formelle

Nous proposons une approche similaire à celle présentée en [7]. De nouvelles connaissances sont intégrées dans le système BYBLOS sur le modèle de la combinaison des processus acoustiques et linguistiques dans le cadre d'une stratégie de décodage asynchrone de type N-Best. Pour notre part, nous penchons pour la fusion de techniques et de connaissances hétérogènes mais dans une stratégie synchrone.

Traditionnellement, l'approche statistique du problème de la reconnaissance de la parole est présentée de la façon suivante [8] :

$$w = \underset{w}{\text{Argmax}}\ P(w/x) = \underset{w}{\text{Argmax}}\ P(x/w).P(w) \quad (1)$$

Avec :

- P(x/w) qui représente la probabilité d'émission de l'observation x pour une phrase w donnée.
- P(w) qui représente la probabilité *a priori* d'une séquence w.

C'est sur le premier terme que nous allons faire porter la fusion des modèles acoustiques. Soit :

$$P(x/w)=\alpha_1.P_1(x/w)+\alpha_2.P_2(x/w)+...+\alpha_i.P_i(x/w) \quad (2)$$

Avec :

$$\alpha_1+\alpha_2+...+\alpha_i = 1 \text{ et } 0 \leq \alpha_k \leq 1 \ ; \ \forall k \leq i$$

Nous avons donc la formule (3) :

$$w=\underset{w}{\text{Argmax}}(\alpha_1.P_1(x/w)+\alpha_2.P_2(x/w)+...+\alpha_i.P_i(x/w)).P(w)$$

$P_k(x/w)$ représentant la probabilité d'émission de l'observation x pour une phrase w donnée et pour une source k de connaissances. Ou plutôt,

un processus k comme par exemple une chaîne de Markov cachée, un réseau de neurones et un ensemble de règles valuées [9].

3.2. Application de la fusion de connaissances hétérogènes

Nous proposons maintenant une application de cette idée dans le cadre de la stratégie de décodage synchrone dite *Modèle de Propagation de Jeton* (ou *MPJ*) [6].

Le *MPJ* propose de voir la reconnaissance de la parole continue à travers un processus de décodage qui s'appuie sur le passage de jetons dans un réseau de transitions. La stratégie de décodage du *MPJ* est indépendante des diverses méthodologies (HMM, RNM, DTW, etc.) utilisées pour le calcul des scores affectés aux unités de base [6, 2].

Pour gérer la fusion de techniques hétérogènes comme définie précédemment, il suffit de modifier légèrement le *MPJ* en considérant le calcul des scores des unités de base comme la somme pondérée de scores issus des différentes méthodologies appliquées à ces mêmes unités (cf. figure 1 et formule 2).

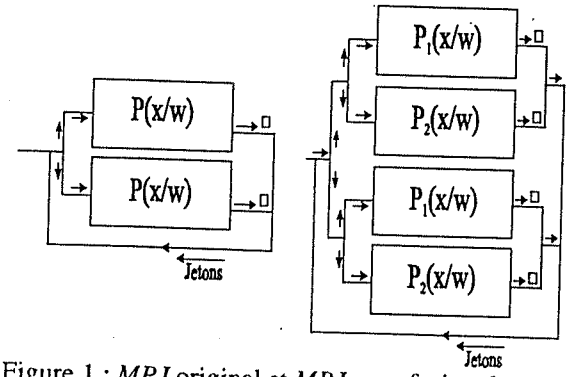


Figure 1 : *MPJ* original et *MPJ* avec fusion de deux types de connaissances hétérogènes.

4. CONCLUSION

Plutôt que la réalisation d'un système de reconnaissance de la parole, nous étudions les moyens de combiner différentes techniques et sources de connaissances au travers des stratégies de décodage. La mise au point d'un système classique de reconnaissance, l'étude de "l'état de l'art" et la recherche d'alternatives nous permettent de guider notre travail.

Mais, avant de pouvoir véritablement évaluer nos approches, nous devons impérativement améliorer le système de base. Dans ce cadre, notre travail porte actuellement sur plusieurs points :

- Intégration de modèles de langage (trigramme, triclasses ou trilemmes).
- La modification des modules de phonétisation afin de générer automatiquement les variantes phonologiques.
- Une augmentation importante des corpus.
- Le choix d'unités acoustiques de bases plus performantes comme des triphones.

Parallèlement, pour répondre aux impératifs de l'hybridation proposée, il nous faut encore mettre au point un module de RNM donnant en sortie des scores probabilistes et capable de gérer la dimension temporelle.

BIBLIOGRAPHIE

[1] H. Bourlard, "Towards increasing speech recognition error rates", *Eurospeech'95*, Madrid, Vol. 2, pp. 883-894, 18-21 septembre 1995.

[2] S. Igounet, "Stratégies de décodage pour la parole continue : fondements, exemples et perspectives", *Rencontres Jeunes Chercheurs en Parole*, ENST Paris, 17-18 novembre 1995.

[3] S.J. Young, P.C.Woodland, "HTK : Hidden Markov Model Toolkit V1.5", *Cambridge University Engineering Departement Speech Group and Entropic Research Laboratories Inc*, 1993.

[4] F. Béchet, S. Derderian, M. El-Bèze, "Conversion graphèmes-phonèmes automatique : le système GRIPHON", *IA 95 - Génie Linguistique*, 15èmes Journées Internationales, Montpellier, juin 1995.

[5] A-M. Derouault, "Context-dependent phonetic Markov Models for Speech Recognition", *NATO Advanced Institute on Pattern Recognition*, pp. 171-175, 5-18 juillet 1987.

[6] S.J. Young, N.H. Russel, J.H.S. Thornton, "Token Passing : a Simple Conceptual Model for Connected Speech Recognition Systems", *Cambridge University Engineering Departement*, rapport technique, 1989.

[7] R. Schwartz, S. Austin, S. Kubala, J. Makhoul, L. Nguyen, P. Placeway, G. Zavaliagkos, "New uses for the N-Best sentence hypotheses within the BYBLOS speech recognition system", *Proc. ICASSP*, pp.11-14, 1992.

[8] D. Jouvet, "Modèle de Markov pour la reconnaissance de la parole", *Ecole thématique : Fondements et Perspectives en Traitement Automatique de la Parole*, Marseille, pp. 99-108, 17-25 juillet 1995.

[9] P. Gilles, "Décodage phonétique de la parole et adaptation au locuteur", *Thèse de doctorat en Informatique*, Université d'Avignon et des Pays de Vaucluse, 1993.

Reconnaissance de la parole continue par le modèle STM polynomial

C. Cerisara, Y. Gong et J.-P. Haton
CRIN – CNRS & INRIA Lorraine, BP 239, F-54506 Vandoeuvre-les-Nancy, France

{gong, jph}loria.fr

RESUME :

Stochastic trajectory models (STM) have been proposed for speech recognition in order to model the correlation between successive observation vectors, assumed to be independent in hidden Markov models (HMM). RFIA group has been developing since recent years STMs, which are used as the base of the VINICS continuous speech recognition system. We present in this paper an extension of STM which models the trajectory evolution of vectors of a phonetic segment by a mixture of polynomial functions of time, allowing an explicit exploitation of correlations between vectors. We first introduce the statistic background of the model and then present preliminary experimental results obtained on a 2000-word continuous speech recognition task.

1 Introduction

Les vecteurs acoustiques successifs dans une séquence d'observation de parole sont par nature corrélés. Or, la plupart des modèles acoustiques destinés à la reconnaissance de la parole ignorent cette corrélation. Par exemple, les modèles de Markov de premier ordre (HMM) supposent que, d'une part, les vecteurs d'observations à l'intérieur d'un segment sont indépendants et générés par la même fonction de densité de probabilité (pdf), et d'autre part, la pdf de chaque segment est indépendante des segments voisins. Par conséquent, l'observation d'une unité sonore est supposée constituée d'une concaténation de segments stationnaires de durée variable.

Les hypothèses d'indépendance ne permettent pas de modéliser les corrélations et peuvent limiter le pouvoir discriminant et la capacité d'apprentissage des HMM pour des contextes phonétiques complexes. Dans ce cadre, de nombreux travaux ont été menés pour tenir compte de la corrélation intra-segment. Les modèles stochastiques de trajectoire (STM) tels que développés notamment par l'équipe RFIA exploitent explicitement cette corrélation, en introduisant un mélange de pdf sur une séquence d'états (Gong & Haton, 1994) où un état correspond à un segment stationnaire de HMM. Cependant, pour une trajectoire donnée, les vecteurs d'observation associés à des états différents sont encore supposés indépendants. Ce problème a été abordé en modélisant l'observation à chaque état comme la somme d'un processus AR non-observable et un bruit spécifique à l'état (Afify, Gong, & Haton, 1996). Le modèle a amélioré sensiblement le taux de reconnaissance, ce qui montre l'importance de modéliser la corrélation, mais il est coûteux en temps de calcul à cause des filtres de Kalman utilisés. Cet article présente un autre modèle de corrélation dans le cadre de STM, dans lequel on suppose que les vecteurs d'observation sont générés aléatoirement par des fonctions polynomiales du temps à paramètres inconnus, avec en plus des sources de bruit. L'estimation des paramètres est obtenue par l'algorithme EM *(expectation-maximisation)*.

Pour les HMM (Deng, 1992), une unique fonction polynomiale est associée à chaque état pour modéliser la trajectoire. Pour le modèle segmental (Gish & Ng, 1993), une mélange de pdfs est définie sur les cofficients polynomiaux. Notre modèle est fondé sur des mélanges et n'a pas besoin de calculer les coefficients polynomiaux pendant la reconnaissance.

2 STM polynomial

2.1 Modèle de base

Nous présentons le principe du modèle stochastique de trajectoire dépendant du temps. Soit $\mathbb{P}$ l'ensemble des classes (phonèmes) à modéliser. Chaque classe $s \in \mathbb{P}$ est associée à un ensemble de générateurs de trajectoires G_s, chacun correspondant à une fonction générique de segments accoustiques avec paramètres inconnus :

$$\forall k \in G_s, \quad f_k^s(i) \in R^D \tag{1}$$

où D est la dimension des vecteurs d'observation, avec la probabilité :

$$\alpha_k^s \triangleq Pr(k|s), \ \forall k \in G_s \tag{2}$$

$$\sum_{k \in G_s} \alpha_k^s = 1 \tag{3}$$

Soit un échantillon d'un segment de Q vecteurs accoustiques :

$$O \triangleq (o_1, o_2, \ldots o_i, \ldots, o_Q), \ \forall i, \ o_i \in R^D \tag{4}$$

Cet échantillon est généré par le générateur k et un vecteur de bruit indépendant spécifique à k :

$$o_i = f_k(i) + n_{i,k} \tag{5}$$

Chaque vecteur de bruit suit une loi gaussienne de moyenne nulle et de matrice de covariance $\Sigma_{i,k}$ de dimension $D \times D$. Connaissant un générateur k et un phonème s, la fonction

de densité de probabilité d'une suite de bruits de longueur Q est :

$$p(n_1, n_2, \ldots n_Q|k, s) = \prod_{i=1}^{Q} p(n_i|k, s) \quad (6)$$

La pdf d'une suite de bruits pour la classe s est alors représentée par un mélange de générateurs :

$$p(n_1, n_2, \ldots n_Q|s) = \sum_{k \in G_s} \alpha_k^s p(n_1, n_2, \ldots n_Q|k, s) \quad (7)$$

Notons que :

$$\begin{aligned} p_N(n_i|k, s) &= N(n_i; 0, \Sigma_{i,k}^s) \\ &= N((o_i - f_k^s(i)); 0, \Sigma_{i,k}^s) \\ &= N(o_i; f_k^s(i), \Sigma_{i,k}^s) \\ &= p_O(o_i|k, s) \end{aligned} \quad (8)$$

où N est la loi normale. Nous supposons que la fonction génératrice f est un polynôme de degré P_k^s :

$$f_k^s(i) = \sum_{j=0}^{P_k^s} c_{k,j}^s \cdot i^j \quad (9)$$

où $c_{k,j}^s \in R^D$ est le vecteur-coefficient d'ordre j pour la composante k du mélange générant s.

Nous représentons les paramètres introduits sous la forme compacte :

$$\lambda \triangleq \{\lambda^s \triangleq \{\alpha_k^s, c_{k,j}^s, \Sigma_{i,k}^s\}\} \quad (10)$$

Dans le paragraphe suivant, nous dérivons les paramètres optimaux en utilisant le critère de maximum de vraisemblance (ML).

2.2 Estimation des paramètres selon EM

2.2.1 Algorithme EM

L'algorithme EM (Dempster, Laird, & Rubin, 1977) permet de trouver une estimation au sens du maximum de vraisemblance de λ. EM calcule itérativement (étapes E et M) :

$$Q(\lambda|\lambda') \triangleq E\{\log p(\mathbf{X}, \mathbf{Y}|\lambda)|\mathbf{X}, \lambda'\}, \lambda' = \underset{\lambda}{\operatorname{argmax}}\, Q(\lambda|\lambda') \quad (11)$$

où X est observable, e.g. les observations d'une trajectoire, Y n'est pas observable, e.g. le générateur de la trajectoire. Soit $\mathcal{O}^s \triangleq \{O_1^s, O_2^s, \ldots O_{M_s}^s\}$ l'ensemble de M_s des observations d'apprentissage associées au phonème s. Appliquons Eq-11 à Eq-7 en tenant compte de Eq-8 :

$$Q(\lambda|\lambda') = \sum_{s \in \mathbb{P}} \sum_{O \in \mathcal{O}^s} \sum_{k \in G_s} \log p(O, k|\lambda) Pr(k|O, \lambda') \quad (12)$$

où $p(O, k|\lambda) = p(O|k, \lambda) Pr(k|\lambda)$,

$$Pr(k|O, \lambda') = \frac{p(O|k, \lambda') Pr(k|\lambda')}{\sum_{k \in G_s} p(O|k, \lambda') Pr(k|\lambda')} \quad (13)$$

$$p(O|k, \lambda) = \prod_{i=1}^{Q} N(o_i; f_k^s(i), \Sigma_{i,k}^s)$$

Eq-12 se simplifie en : $Q(\lambda|\lambda') = \sum_{s,O,k} P(k|O, \lambda') \cdot$

$$\sum_{i=1}^{Q} \log N(o_i; f_k^s(i), \Sigma_{i,k}^s) + P(k|O, \lambda') \log \alpha_k^s \quad (14)$$

Nous maximisons la fonction Q par rapport à chaque paramètre dans λ.

2.2.2 Probabilités a priori des générateurs

Pour obtenir l'estimation de la probabilité a priori α_k^s, nous appliquons la méthode de Lagrange à l'équation Eq-14 avec la contrainte de l'équation Eq-3. Nous obtenons :

$$\alpha_k^s = \frac{1}{M_s} \sum_{O \in \mathcal{O}^s} P(k|O, \lambda'), \qquad k \in G_s \quad (15)$$

2.2.3 Fonctions génériques polynomiales

Dériver directement Eq-14 selon les coefficients de la fonction génératrice fournirait un ensemble d'équations dépendant à la fois de $c_{k,j}^s$ et de $\Sigma_{i,k}^s$. La solution de telles équations n'est pas aisée à cause du terme $\Sigma_{i,k}^s$, qui doit être simplifié. Les expériences préalablement menées par l'équipe montrent que la variance est plus grande aux extrémités d'un segment de parole qu'en son centre. Nous supposons dès lors que :

$$\Sigma_{i,k}^s \triangleq \omega_{i,k}^s \Sigma_k^s \quad (16)$$

où $\omega_{i,k}^s$ est strictement positif. A l'aide de Eq-16, nous pouvons dériver Eq-14 par rapport aux coefficients de la fonction génératrice. Nous obtenons :
$\sum_{p=0}^{P_k^s} \sum_{O \in \mathcal{O}^s} \sum_{i=1}^{Q} \eta^s(i, k, O, p+j) \cdot c_{k,p}^s =$

$$\sum_{O \in \mathcal{O}^s} \sum_{i=1}^{Q} \eta^s(i, k, O, j) \cdot o_i \quad (17)$$

avec

$$\eta^s(i, k, O, m) \triangleq P(k|O, \lambda') \frac{i^m}{\omega_{i,k}^s} \quad (18)$$

Eq-17 décrit, pour chaque composante k du mélange, un système linéaire à $P_k^s + 1$ inconnues, chaque variable étant un vecteur de dimension R^D. Ceci peut s'écrire sous la forme compacte :

$$\forall k \in G_s, \quad \mathbf{A}_k^s \mathbf{c}_k^s = \mathbf{b}_k^s \quad (19)$$

$\mathbf{A}_k^s$ est une matrice de dimension $(P_k^s + 1) \times (P_k^s + 1)$

$$\mathbf{A}_k^s \triangleq \left[a_{k,(p,j)}^s\right] \quad (20)$$

avec $a_{k,(p,j)}^s \triangleq \sum_{O \in \mathcal{O}^s} \sum_{i=1}^{Q} \eta^s(i, k, O, p+j)$. $\mathbf{b}_k^s$ est un vecteur de dimension $(P_k^s + 1)$:

$$\mathbf{b}_k^s \triangleq [b_{k,0}^s, \ldots b_{k,j}^s, \ldots b_{k,P_k^s}^s]^t \quad (21)$$

où $b^s_{k,j} \triangleq \sum_{O\in\mathcal{O}^s}\sum_{i=1}^{Q}\eta^s(i,k,O,j)\cdot o_i$. De même, $\mathbf{c}^s_k$ est un vecteur de dimension (P^s_k+1) :

$$\mathbf{c}^s_k \triangleq [c^s_{k,0},\ldots c^s_{k,j},\ldots c^s_{k,P^s_k}]^t \quad (22)$$

Ainsi, $\mathbf{c}^s_k$ peut être calculé à l'aide d'une méthode adéquate de résolution de systèmes linéaires.

2.2.4 Matrice de covariance

Pour calculer Σ^s_k, on dérive Eq-14 par rapport à $(\Sigma^s_k)^{-1}$ et on cherche la racine de l'équation obtenue. Soit :

$$\Sigma^s_k = \frac{\sum_{O\in\mathcal{O}^s} p(k|O,\lambda')\sum_{i=1}^{Q}\frac{(o_i-f^s_k(i))(o_i-f^s_k(i))^t}{\omega^s_{i,k}}}{Q\sum_{O\in\mathcal{O}^s}p(k|O,\lambda')} \quad (23)$$

où $k\in G_s$. En option, si les matrices de covariance de tous les mélanges sont liées, i.e. : $\Sigma^s_k \rightarrow \Sigma^s$, Eq-23 devient :

$$\Sigma^s = \frac{\sum_{O,k} p(k|O,\lambda')\sum_{i=1}^{Q}\frac{(o_i-f^s_k(i))(o_i-f^s_k(i))^t}{\omega^s_i}}{Q\cdot M_s} \quad (24)$$

2.2.5 Facteur de covariance

Le facteur de la matrice de covariance $\omega^s_{i,k}$ peut également être calculé à l'aide de l'algorithme EM :

$$\omega^s_{i,k} = \frac{\sum_{O\in\mathcal{O}^s} p(k|O,\lambda')(o_i-f^s_k(i))^t{\Sigma^s_k}^{-1}(o_i-f^s_k(i))}{D\sum_{O\in\mathcal{O}^s}p(k|O,\lambda')} \quad (25)$$

et

$$\omega^s_i = \frac{1}{DM_s}\sum_{O,k} p(k|O,\lambda')(o_i-f^s_k(i))^t{\Sigma^s}^{-1}(o_i-f^s_k(i)) \quad (26)$$

pour les états liés et, pour les états non liés :

$$\omega^s_i = \frac{1}{DM_s}\sum_{O,k} p(k|O,\lambda')(o_i-f^s_k(i))^t{\Sigma^s_k}^{-1}(o_i-f^s_k(i)) \quad (27)$$

Nous observons que l'expression de $\omega^s_{i,k}$ dépend de $f^s_k(i)$ et de Σ^s_k. Il n'est pas facile de donner une solution analytique à cet ensemble d'équations en séparant les trois inconnues. Toutefois, deux techniques peuvent être utilisées pour calculer $\omega^s_{i,k}$. La première, que nous utilisons au § 3.2.2, consiste à substituer Σ^s_k par $\Sigma^{s\prime}_k$ obtenu à l'itération EM précédente. La seconde est d'utiliser d'autres critères d'apprentissage tels que l'information mutuelle maximale (Duda & Hart, 1973; Li, Haton, & Gong, 1995), ou l'erreur minimale (Juang & Katagiri, 1992; Chou, Juang, & Lee, 1992).

3 Expérimentations

3.1 Conditions expérimentales

La tâche traitée consiste à reconnaître des rapports oraux d'inspection de centrale nucléaire dans le cadre d'une étude avec le CEA-cadarache. Une grammaire de 2010 mots de perplexité bi-grammes 48 est utilisée. L'apprentissage, dépendant du locuteur, a été réalisé sur 80 phrases (6 minutes) et l'évaluation sur 241 autres phrases (1482 mots). L'apprentissage est en mode indépendant de la tâche, c'est-à-dire que peu de mots à reconnaître sont prononcés pendant l'apprentissage.

32 modèles phonémiques à 5 états (Q=5) sont introduits avec de 1 à 8 trajectoires par mélange par phonème, selon le nombre d'observations de phonèmes disponibles. Dans les modèles STM de base, les matrices de covariance sont supposées diagnonales et définies sur 13 coefficients mel-cepstaux. En moyenne, 600 lois gaussiennes multidimensionelles par locuteur sont générées.

Nos expériences ont montré que la variance des vecteurs acoustiques appartenant à des générateurs différents est plus faible au milieu d'un segment phonétique qu'aux extrémités. De ce fait, dans cette étude, les paramètres des états centraux ($i=2$) sont liés.

Pendant l'apprentissage par EM, les générateurs de trajectoire dont le nombre de trajectoires du corpus d'apprentissage associées est inférieur à MINOBS sont éliminés, à cause de la non fiabilité de l'estimation. Expériementalement, MINOBS a été fixé à 6 pour le STM de base.

Le degré de la fonction génératrice est fixé à $Q-1$, puisqu'un unique polynôme de degré $Q-1$ est entièrement déterminé par $Q=5$ points. Cependant, ce degré pourrait être trop élevé pour représenter les trajectoires d'une manière lissée.

Nous appellerons dans la suite *test A* le test de reconnaissance phonétique sur le corpus d'apprentissage, et *test B* le test de reconnaissance portant sur des phrases n'ayant pas servi à l'apprentissage.

3.2 Résultats préliminaires

3.2.1 Observations générales

Sur un même corpus de données, nous avons calculé un échantillon des moyennes et des covariances pour STM de base et STM polynomial. Les moyennes étaient très proches dans les deux cas. Les covariances étaient également proches, mais des différences nettement plus visibles. Compte tenu des différences de formalismes et des approximations réalisées sur les covariances, cette différence parait normale.

Tab-1 montre les taux de reconnaissance obtenus avec le STM de base, utilisé pour comparaison.

test	STM MINOBS=6
test A	85.6%
test B	99.5%

TAB. 1 – *Taux de reconnaissance avec le STM de base*

3.2.2 Facteur de covariance

Différentes manières de calculer $\omega^s_{i,k}$ ont été testées :

- Valeur fixée arbitrairement. La courbe des $\omega^s_{i,k}$ fournissant le meilleur résultat est : $\forall s,\forall k,\quad \omega^s_{0,k} =$

$1.3, \omega^s_{1,k} = 1.1, \omega^s_{2,k} = 1.0, \omega^s_{3,k} = 1.2, \omega^s_{4,k} = 1.5$; Cette courbe correspond approximativement aux variances par état, normalisées sur celle de l'état central.

- Valeur calculée à l'aide des formules Eq-25 à Eq-27. Cette méthode semble a priori la meilleure, mais il faut rappeler qu'il n'est pas possible d'utiliser la matrice de covariance *actuelle*, car celle-ci dépend elle-même de $\omega^s_{i,k}$. Nous utilisons par conséquent la matrice de covariance calculée lors de l'itération précédente de l'algorithme EM, en espérant que celle-ci est suffisamment proche de la matrice théorique.

Les résultats (Tab-2) sont proches avec une petite différence sur la reconnaissance phonétique sur le corpus d'apprentissage.

test	fixée	estimée
test A	86.2	86.3
test B	99.3	99.3

TAB. 2 – *Taux de reconnaissance avec le STM polynomial en fonction de choix de calcul de* $\omega^s_{i,k}$

3.2.3 Nombre minimum de trajectoires

Dans le STM polynomial, les états sur une trajectoire entière sont liés par la fonction polynomiale. Par conséquent, MINOBS pourrait être diminué grâce à l'accroissement de la robustesse. Tab-3 confirme cette hypothèse.

test	MINOBS=6	MINOBS=4
test A	84.5%	84.9%
test B	98.5%	99.2%

TAB. 3 – *Taux de reconnaissance avec le STM polynomial en fonction du nombre minimal d'observations par générateur*

3.2.4 Matrices de covariance liées

Un autre test a été réalisé afin de mesurer l'importance réelle de la matrice de covariance dans la discrimination. Ce test consiste à ne construire qu'une seule matrice de covariance par symbole, indépendamment des mélanges (Eq-24). Cela revient à supposer que seules les moyennes ont de l'importance pour la reconnaissance, les matrices de covariance pouvant être grossièrement approchées (Ney, 1993). Tab-4

test	Σ^s_k	Σ^s
test A	85.3%	83.4%
test B	99.2%	99.1%

TAB. 4 – *Taux de reconnaissance avec le STM polynomial en fonction du type de matrices de covariance*

montre que les résultats obtenus avec les matrices de covariance liées sont légèrement inférieurs aux résultats obtenus en calculant plus précisément la matrice de covariance, ce qui tendrait à confirmer la théorie selon laquelle la matrice de covariance n'est pas importante dans la discrimination.

4 Conclusion

Nous avons présenté dans cet article un modèle de corrélation pour le modèle stochastique de trajectoire (STM) que nous avons mis au point dans notre équipe.

Dans la nouvelle modélisation, au lieu d'être indépendants, les états successifs d'une trajectoire sont représentés comme des échantillons d'une courbe polynomiale. Par conséquent, l'estimation des paramètres, en particulier les moyennes, de la trajectoire devient plus fiable.

Les résultats expérimentaux préliminaires obtenus montrent que les performances sont équivalentes à celle du modèle STM initial. Ces premiers résultats sont très encourageants et devraient nous permettre d'améliorer à terme les taux de reconnaissance en parole continue.

Pour évaluer l'apport de la modélisation polynomiale, il est nécessaire d'effectuer des tests en fonction du degré du polynôme. Une extension intéressantes de la méthode proposée consisterait à exploiter les covariances associées aux différents états d'une trajectoire pour améliorer la discrimination entre des trajectoires confondues par le modèle STM de base.

Références

Afify, M., Gong, Y., & Haton, J.-P. (1996). Estimation of mixtures of stochastic dynamic trajectories: Application to continuous speech recognition. *Computer, Speech and Language, 10*, 23–36.

Chou, W., Juang, B. H., & Lee, C. H. (1992). Segmental GPD training of HMM based speech recognizer. In *Proc. of IEEE Internat. Conf. on Acoustics, Speech and Signal Processing*, Vol. I, p. 473.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, 39*(1), 1–38.

Deng, L. (1992). A generalized hidden Markov model with state-conditioned trend functions of time for the speech signal. *Signal Processing, 27*(1), 65–78.

Duda, R. O., & Hart, P. E. (1973). *Pattern Classification and Scene Analysis*. John Wiley & Sons, New York.

Gish, H., & Ng, K. (1993). A segmental speech model with applications to word spotting. In *Proc. of IEEE Internat. Conf. on Acoustics, Speech and Signal Processing*, Vol. II, pp. 447–450.

Gong, Y., & Haton, J.-P. (1994). Stochastic trajectory modeling for speech recognition. In *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Vol. I, pp. 57–60 Adelaide, Australia.

Juang, B.-H., & Katagiri, S. (1992). Discriminative learning for minimum error classification. *IEEE Trans. on Signal Processing, 40*(12), 3043–3054.

Li, H. Z., Haton, J.-P., & Gong, Y. (1995). On the map learning of Gaussian mixture for speaker models. In *Proceedings of European Conference on Speech Technology*, pp. 363–366 Madrid.

Ney, H. (1993). Modeling and search in continuous speech recognition. In *Proceedings of European Conference on Speech Technology*, Vol. 1, pp. 491–498 Berlin.

RECONNAISSANCE DE PAROLE EN MILIEU BRUITÉ : CONTRIBUTION À LA ROBUSTESSE DES SYSTÈMES

Jean-Baptiste PUEL
IRIT - Université Paul Sabatier - 118, route de Narbonne
31062 Toulouse cedex - France
puel@irit.fr

ABSTRACT

Speaker independent recognition systems designed in adverse conditions (as telephonic networks) deal with small vocabularies and offer good results using Hidden Markov Models (HMM). In order to obtain good recognition performances, the recording conditions must be as close as possible from the real using conditions. In some case, it should be restricting, and sometimes impossible to record the learning corpus in the real recognition conditions. We propose two kinds of methods to release this constraint and to enhance speech recognition rates : a noise subtraction preprocessing method and a model architecture modification one. We evaluate these approachs with a segmental HMM recognition system.

1. INTRODUCTION

La reconnaissance de petits vocabulaires, indépendamment du locuteur et sous la forme de mots isolés, semble être un problème résolu depuis la généralisation des systèmes à base de Modèles de Markov Cachés (MMC).
De tels systèmes offrent des solutions réalistes pour la création de serveurs vocaux interactifs, de systèmes à commande vocale comme les composeurs téléphoniques. Cependant, leur point faible est le manque de robustesse au changement d'environnement: de légères différences entre les conditions d'apprentissage et les conditions réelles d'utilisation suffisent souvent à dégrader leurs performances.
En condition réelle d'utilisation, nombreuses sont les causes de variabilité pouvant générer un signal différent de celui utilisé lors de l'apprentissage : l'environnement sonore, le procédé et le matériel d'acquisition du signal, le locuteur lui même. Les Modèles de Markov Cachés, bien que robustes, ne permettent pas de passer outre une telle variabilité; dans ce contexte, nous avons mis en oeuvre deux grands types de méthodes pour améliorer les performances des systèmes de reconnaissance : des techniques de débruitage et des architectures ou structures de systèmes robustes.

2. MÉTHODES DE DÉBRUITAGE

On considère que le signal de parole bruité résulte de la modification du signal de parole "propre" par divers filtres : additifs pour le bruit ambiant, convolutifs pour l'enregistrement et la transmission du signal, autres (bruit impulsionnel). La solution générale de débruitage que nous proposons consiste à évaluer des paramètres représentatifs du bruit pour chacun de ces filtres. Pour cela, nous avons mis au point plusieurs modules de prétraitement : ces modules sont utilisés aussi bien lors de l'apprentissage des modèles que lors de la reconnaissance.

2.1. La segmentation automatique

L'algorithme de segmentation utilisé est la méthode de "Divergence Forward-Backward" [André-Obrecht 88], qui localise les zones quasi-stationnaires de signal. Ce premier module du prétraitement est effectué systématiquement, en effet, l'information apportée par la position des ruptures dans le signal est utilisée par tous les autres modules. On considère que le signal est représenté par une chaîne d'unités homogènes, chacune d'entre elles représentée par un modèle AR. La méthode consiste à détecter les changements dans les paramètres du modèle. Les segments obtenus peuvent être rangés dans trois catégories : des segments stationnaires correspondant aux parties stables du signal, des segments de transition dans lesquels on trouve une structure formantique et dont le comportement reste monotone, des segments courts (de l'ordre de 10 ms) qui correspondent à des changements articulatoires rapides, comme l'explosion d'une plosive. L'expérience a montré que cet algorithme de segmentation détecte toutes les frontières bruit/parole, mais sans toutefois les identifier.

2.2. La détection bruit/parole

Le module de segmentation fournit une liste de frontières correspondant à des modifications spectrales du signal. Cependant, on ne sait pas a priori pour chacun des segments s'il est situé dans une zone de parole ou dans une zone de bruit. Dans la littérature, de nombreux détecteurs bruit/parole sont proposés, mais rares sont ceux qui tirent parti d'un algorithme de segmentation. Diverses techniques sont fondées sur la reconnaissance des formes, comme [Lamel 81], d'autres sur l'utilisation de coefficients acoustiques particuliers comme le taux de passage par zéro, l'énergie,... [Mak 92]. L'approche que nous avons retenue est basée sur le fait que même si l'énergie

du signal n'est pas un paramètre robuste en milieu bruité, ses maxima d'amplitude correspondent toujours à des noyaux vocaliques [Puel 94]. Ce traitement se déroule en deux phases :

2.2.1 L'étiquetage statique

Dans un premier temps, l'abscisse curviligne $s(t)$ du signal de parole $y(t)$, où t est l'indice des échantillons, est calculée. Soit la fonction :

$$S(n) = s(nL) - s((n-1)L)$$

où L est un nombre d'échantillons fixé (une trame). $S(n)$ représente une valeur moyenne de la "longueur de la courbe" par unité de temps. En supposant que le bruit est stationnaire pour chaque segment, la fonction S varie peu dans les zones de bruit, croît très sensiblement avec l'apparition de la parole pour décroître lors de sa disparition. Ce sont les moyennes $\overline{S_i}$ et écart-types $\sigma(S_i)$ de S pour le segment i qui vont représenter notre indicateur de bruit ou parole. Deux seuils sont utilisés : λ_1 et λ_2, calculés automatiquement sur les trames b du signal supposées n'être que du bruit (et où S présente ses minima).

$$\lambda_1 = \overline{S_b} + \sigma(S_b) \text{ et } \lambda_2 = n \times \lambda_1$$

où n est un seuil permettant de discriminer le niveau moyen de l'abscisse curviligne entre le bruit et la parole bruitée. Sur les corpus de signal RTC et Itineris, l'ordre de grandeur pour n est 3.
L'étiquetage statique consiste à comparer les moyennes et écart-types de S par segment à ces seuils.

2.2.2 La coordination temporelle

Nous appliquons les règles suivantes à chacun des segments : un segment isolé de parole est classé "bruit", un court segment de bruit entre deux segments de parole sera classé "parole" s'il ne dépasse pas 80 ms (tenue de plosive). Enfin, un traitement d'exceptions est prévu pour quelques situations particulières (SNR très faible, présence d'artefacts juste avant ou après le mot prononcé) où l'étiquetage statique se déroule anormalement. Typiquement, si moins de trois segments de parole sont détectés, les seuils sont réévalués et la détection est élargie autour du segment où l'activité "parole" est la plus vraisemblable. A l'issue de cette étape, nous disposons de l'étiquetage définitif des segments, et par là même, de la position de la parole dans le signal.

2.3. Le débruitage

Le point commun de la plupart des techniques de débruitage est d'opérer la soustraction du bruit estimé dans le domaine spectral, permettant de résoudre pour une grande part le problème du bruit additif [Boll 79].
Nous avons implémenté l'algorithme de Soustraction Spectrale Linéaire du bruit (SSL).
Cette méthode consiste à retrancher une estimation de la densité spectrale de puissance du bruit de la densité spectrale de puissance du signal bruité.

La densité spectrale de puissance du bruit est calculée dans les zones détectées comme du bruit seul, dans la plus longue portion de bruit précédant ou suivant le mot à débruiter. On considère que le signal $x(n)$ est l'addition de la parole $s(n)$ et d'un bruit aléatoire non corrélé $b(n)$, stationnaires à court terme.

$$x(n) = s(n) + b(n) \quad (1)$$

Dans le domaine spectral

$$\Gamma_x(\omega) = \Gamma_s(\omega) + \Gamma_b(\omega) \quad (2)$$

où $\Gamma_x(\omega)$ représente la densité spectrale de puissance à court terme de $x(n)$. En reconnaissance de parole, la phase $\phi(\omega)$ apporte peu d'informations, aussi peut-on la négliger. On calcule le signal débruité par :

$$\hat{S}(\omega) = |X(\omega)| - |\hat{B}(\omega)| \quad (3)$$

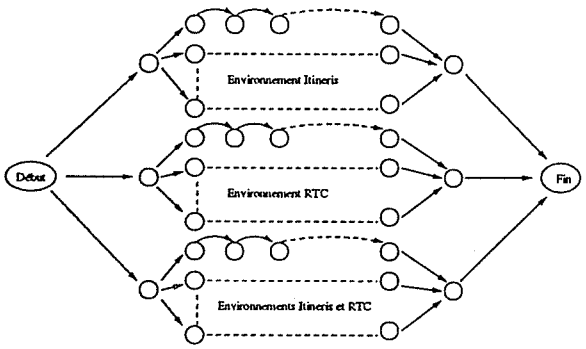
Toute la difficulté réside dans l'estimation du bruit $\hat{B}(\omega)$. C'est à ce stade que nous utilisons les résultats du détecteur bruit/parole : la plage de bruit la plus longue détectée juste avant ou après le mot prononcé nous permet de mettre à jour l'estimation du bruit courant. Le spectre du bruit est estimé sur le plus possible de fenêtres de 32 ms, avec un recouvrement de 16 ms. Le vecteur représentatif conservé est la moyenne du spectre sur toutes ces fenêtres. Un des risques de cette méthode est d'opérer une soustraction trop brutale dans les zones de faible SNR, où le signal de parole risque d'être confondu avec le bruit, aussi lorsque le rapport signal/bruit est faible (en dessous de $5dB$), nous appliquons un coefficient de plus en plus petit. La priorité étant, en effet, de privilégier l'intégrité du signal de parole sur l'efficacité du débruitage.

3. STRUCTURES DE MODÈLES ROBUSTES

Au lieu de chercher à supprimer le bruit additif comme dans les méthodes de débruitage, la conception de structures de modèles robustes vise à tenir compte de la présence du bruit dans la structure des modèles soit en modélisant simultanément parole et bruit par deux Modèles de Markov ; soit en construisant un modèle de parole bruitée à partir d'un modèle de parole et d'un modèle de bruit : combinaison parallèle de modèles de [Gales et Young 92]. Une autre méthode - celle que nous avons retenue - consiste à entraîner les modèles avec de la parole bruitée, autant que possible dans plusieurs environnements de bruit différents, pour offrir la plus grande variété possible lors de l'apprentissage. Reste à concevoir des modèles capables de porter ce supplément d'information : nous proposons des *Multi-Modèles* ou des modèles à *Multi-Gaussiennes*. Ces deux méthodes sont comparées dans [Jouvet 94], qui montre qu'une augmentation du nombre de paramètres des modèles est généralement bénéfique pour les performances des systèmes de reconnaissance.

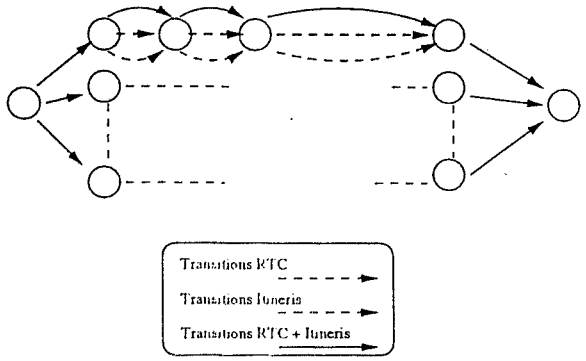
3.1. Les Multi-Modèles

L'idée de cette méthode est de disposer de plusieurs modèles différents pour représenter chaque unité à reconnaître. Un exemple classique consiste à modéliser séparément les locuteurs homme et femme. Deux modèles sont appris séparément sur une partition arbitraire des données (ici hommes et femmes) puis réunis par un état de début et de fin communs. Nous nous proposons d'expérimenter cette méthode en utilisant comme partitions des données d'apprentissage les différents contextes de bruit dont nous pouvons disposer. Le Multi-Modèle réalisé est le suivant: un modèle appris sur un contexte de bruit (Itineris), un modèle appris sur un autre contexte (RTC), un modèle appris sur les deux contextes à la fois. Les trois modèles sont réunis par un état de début et de fin communs.



3.2. Les Multi-Gaussiennes

Afin de rendre plus discriminantes les distributions probabilistes, une approche couramment employée consiste à utiliser plusieurs lois de probabilité.



Chacune des lois est initialisée séparément par l'apprentissage d'un corpus particulier, ensuite de quoi toutes les lois de probabilité du système sont réestimées par un apprentissage complet de tous les corpus. Dans le cadre qui nous intéresse, nous allons initialiser chacune des lois par l'apprentissage d'un corpus correspondant aux données enregistrées dans un contexte particulier de bruit, et ensuite effectuer l'apprentissage global du système en utilisant toutes les données disponibles.

4. EXPÉRIMENTATIONS

Les différentes techniques présentées ont été expérimentées en réalisant différents systèmes de reconnaissance grâce au compilateur de MMC de l'IRIT [Jacob 95]. L'approche retenue consiste à décrire le vocabulaire de l'application par des réseaux de pseudo-diphones. Les modèles acoustiques sont classiques et élémentaires, les lois d'observation sont des gaussiennes simples de matrice de covariance diagonale. Un seul vecteur d'observations est retenu pour chaque segment, il correspond à l'analyse spectrale de la fenêtre centrale du segment. Ce vecteur est composé de : 8 coefficients MFCC et leurs 8 dérivées, l'énergie de la fenêtre, et sa dérivée, la longueur du segment. Les corpus utilisés sont deux corpus de commandes du CNET comprenant 16 mots prononcés par une centaine de locuteurs au travers des réseaux téléphoniques commuté (RTC) et radio-mobile (Itineris). La moitié des données est utilisée pour réaliser l'apprentissage, la seconde moitié pour effectuer les tests.

4.1. Résultats du débruitage

Les résultats de départ obtenus par le système de reconnaissance sont les suivants :

	Réseau RTC	Itineris
Test RTC	7.2 %	16.0 %
Test ITI	10.3 %	7.8 %

Ce tableau présente les taux d'erreurs obtenus par le système segmental de base sur l'ensemble test lorsque le signal ne subit aucun prétraitement, les colonnes présentent le corpus d'apprentissage, les lignes le corpus de test. Bien entendu, les résultats les plus mauvais sont obtenus dans des conditions de reconnaissance croisée (apprentissage sur Itineris, test sur RTC : 16 % d'erreurs).
La méthode de débruitage par SSL et coefficient de pondération présente les résultats suivants :

	Réseau RTC	Itineris
Test RTC	4.6 %	12.6 %
Test ITI	6.6 %	6.2 %

Une amélioration moyenne de 28% est obtenue par cette méthode, sur les deux corpus.

4.2. Résultats de la modification de la structure des modèles

Les *Multi-Modèles* sont la réunion de plusieurs modèles appris séparément, et associés par un état de départ et de fin communs. Nous avons utilisé des MMC entraînés indépendamment sur chaque contexte

de bruit (Itineris, RTC, les deux), réunis et évalués sur les données de test des deux environnements. Comme le montre le tableau suivant, cette méthode améliore sensiblement performances et robustesse.

	Multi-Modèles
Test RTC	4.7 %
Test ITI	5.0 %

Pour tester les modèles à base de *Multi-Gaussiennes*, nous avons créé un système segmental comportant 3 lois par transition. Chaque loi est apprise indépendamment sur un corpus correspondant à un milieu bruité particulier.
L'apprentissage se déroule en deux étapes :
Initialisation des lois : un premier apprentissage est effectué pour la première loi sur le corpus Itineris. Un deuxième apprentissage pour la deuxième loi (sans réestimer la première) sur le corpus RTC. Enfin, un dernier apprentissage pour la troisième loi (sans réestimer les deux autres) sur les deux corpus Itineris et RTC.
Apprentissage global : dans un second temps, le modèle est appris globalement, c'est-à-dire que toutes les lois sont réestimées ensemble pour les exemples proposés, sur toutes les données de tous les environnements.
Cette méthode contribue également à réaliser des systèmes plus robustes.

	Multi-Gaussiennes
Test RTC	3.4 %
Test ITI	4.6 %

Il est possible de combiner les deux structures de réseaux en réalisant des *Multi-Modèles de Multi-Gaussiennes*. Il s'agit de regrouper trois modèles de Multi-Gaussiennes entraînés séparément en un seul Multi-Modèle.
Nous avons donc constitué successivement trois modèles comme au paragraphe précédant, le premier initialisé sur Itineris puis appris sur RTC, inversément pour le deuxième, le dernier initialisé et appris sur les deux environnements à la fois.
Ces trois modèles ont été regroupés avec un état de début et de fin communs, puis évalués sur les deux corpus de test :

	Multi-Mod. et Gauss.
Test RTC	2.6 %
Test ITI	3.6 %

C'est ce modèle qui présente les meilleurs résultats et qui offre la meilleure robustesse : globalement 70% d'erreurs en moins, mais il comporte trois fois plus d'états et neuf fois plus de lois de probabilité que le modèle de base.

5. CONCLUSION

La méthode de débruitage offre une amélioration du taux d'erreur de 30%, au prix de calculs supplémentaires, mais sans changer la taille des modèles. Cette méthode et ses variantes sont particulièrement adaptées aux bruits stationnaires de niveau élevé et limités à certaines bandes de fréquence, comme le bruit d'un moteur d'automobile.
En revanche, pour les contextes de bruit qui nous intéressent : bruit de canal téléphonique, ou radiomobile, il semble que cette méthode atteigne ses limites.
Les méthodes consistant à apprendre les systèmes sur plusieurs environnements (Multi Gaussiennes ou Multi Modèles) présentent une bonne robustesse et d'excellentes performances, pouvant aller jusqu'à supprimer 70% des erreurs, certainement grâce à l'augmentation du nombre de paramètres des modèles.
Cependant, il n'est peut-être pas toujours possible de disposer de suffisamment de données différentes pour réaliser un tel apprentissage. Une perspective intéressante dans ce domaine est l'adaptation des références des systèmes, c'est-à-dire déterminer une transformation permettant de rapprocher les références enregistrées dans un contexte de bruit d'un autre contexte de bruit.

Références

[André-Obrecht 88] R. ANDRÉ-OBRECHT, "A New Statistical Approach for Automatic Segmentation of Continuous Speech Signals", IEEE Trans. on ASSP, vol. 36 pp 26-40, January 1988.

[Boll 79] S.F. BOLL, "Suppression of acoustic noise in speech using spectral subtraction", IEEE Trans. on ASSP, vol. 27 pp 113-120, 1979.

[Gales et Young 92] M.J.F. GALES and S. YOUNG, "An Improved Approach to the Hidden Markov Model Decomposition of Speech and Noise", ICASSP 1992.

[Jacob 95] B. JACOB, "Un Outil Informatique de Gestion de Modèles de Markov Cachés : Expérimentation en Reconnaissance Automatique de la Parole", Thèse de l'Université Paul Sabatier, Toulouse, 1995.

[Jouvet 94] D. JOUVET, M. DAUTREMONT et A. GOSSART, "Comparaison des Multi-Modèles et des Densités Multi-Gaussiennes pour la Reconnaissance de la Parole par Modèles de Markov", JEP 1994.

[Lamel 81] L.F. LAMEL, L.R. RABINER, A.E. ROSENBERG, J.G. WILPON "An Improved Endpoint Detector for Isolated Word Recognition", IEEE Trans. on ASSP, Vol. 29, pp 777-785, August 1981.

[Mak 92] B. MAK, J.C. JUNQUA, B. REAVES, "A robust speech/non-speech detection algorithm using time and frequency-based features", ICASSP 1992.

[Puel 94] J-B. PUEL, R. ANDRÉ-OBRECHT, "Robust Signal Preprocessing for HMM Speech Recognition in Adverse Conditions", ICSLP 1994.

RECONNAISSANCE DE LA PAROLE : VERS L'UTILISABILITÉ

Jean CAELEN, Harouna KABRE, Olivier DELEMAR

Laboratoire CLIPS-IMAG - Domaine universitaire, BP 53 - 38041 Grenoble Cedex 9

Tél. : 76.51.46.27 - Fax : 76.44.66.75 - Email : Jean.Caelen@imag.fr

Résumé

The paper describes a speech recognition software architecture devoted to the usability problem. Robustness is one of the necessary but not sufficient conditions for usability. Usability means stability of the performance in all usage conditions rather than high performance in a limited set of usage conditions. In order to reach this objective, the system must anticipate any changes in the environment and adapt itself to the different usage conditions. A solution to this problem is to connect specialised modules of pre-processing and post-processing to the general recognition engine. These modules interact as a "mirror" with the engine in order to improve its performance and to force it to revise its own knowledge and consequently improve its learning method

1. INTRODUCTION

Les performances des SRAP (Systèmes de Reconnaissance Automatique de la Parole) obtenues en laboratoire sont souvent suffisantes pour conduire à des systèmes potentiellement utilisables. Or les systèmes ne sont pas toujours utilisables malgré leurs performances affichées. L'utilisabilité est donc bien la notion centrale pour un SRAP. Malheureusement, c'est une notion délicate à définir car elle est relative aux utilisateurs concernés et aux domaines d'application. L'utilisabilité d'un système est conditionnée par sa robustesse [Gong 95]. La robustesse en ce sens traduit l'invariance des performances en environnements variables plus que des performances élevées en conditions de fonctionnement aseptisées : c'est cette régularité de comportement du système qu'il faut atteindre. Pour cela il ne s'agit pas d'améliorer uniquement le moteur du SRAP mais de le solidifier vis-à-vis des conditions d'utilisation adverses – d'ailleurs les améliorations ne sont plus guère possibles quand on est déjà à des taux supérieurs à 95% dans les modèles de Markov : l'introduction de contraintes comme les durées n'améliore pas vraiment les performances du système par rapport au coût payé en apprentissage [Delemar, 95] ou en finesse d'ajustement ; on sait aussi que des méthodes mixtes fondées connaissances [Rossi 95] ou fondées réseau neuronaux [Haton 95] ne sont pas non plus la panacée.

Il est bon de remarquer tout d'abord que malgré des performances suffisantes en régime de pointe, les systèmes markoviens ont un mauvais taux de rejet d'une part, et sont fragiles dans des situations qui s'éloignent des situations d'apprentissage d'autre part : il ne suffit pas de prévoir un apprentissage robuste dans des situations variées car nul ne peut prévoir à l'avance les conditions dans lesquelles les systèmes de reconnaissance vont être utilisés et les situations que l'on aura imaginées seront toujours trop limitées voire limitantes. Cela justifie donc que l'on se préoccupe de la robustesse des systèmes si l'on veut les inclure dans des systèmes de dialogue homme-machine.

La robustesse résulte à notre sens de l'adaptabilité et/ou de l'adaptativité du système. Si l'on se tourne vers les êtres humains cela semble en effet être leur force. L'être humain est capable de comprendre des messages dans des milieux très bruyants, à travers des canaux de transmission déformants, pour des timbres voilés, etc. mais contrairement à la machine il sait caractériser son environnement : cela voudrait-il dire qu'il "intègre" l'environnement plus qu'il ne l'élimine ?

Dans un système de reconnaissance on tente généralement de filtrer le bruit, de rehausser le signal, etc., ne serait-ce donc pas l'inverse qu'il faudrait faire ? Ne

faudrait-il donc pas plutôt "reconnaître" les bruits perturbateurs plutôt que d'essayer de les éliminer ? Ne faudrait-il pas aussi prévoir qu'un locuteur peut se déplacer en parlant ? N'est-ce pas aussi l'intérêt d'un système de reconnaissance de pouvoir lui parler sans avoir "la bouche rivée au microphone" ?

Ce sont ces problèmes liés à l'utilisabilité qu'il faut aborder nous semble-t-il maintenant, si l'on veut que les systèmes de reconnaissance soient utiles et utilisés dans des applications. C'est à ce poblème que s'adresse cet article qui propose quelques réflexions et des premiers résultats de mise en œuvre dans le système ECHO (Environnement de Communication parlée Homme-Ordinateur).

2. LE SYSTEME "ECHO"

L'architecture (Fig. 1) que nous proposons s'articule autour d'un moteur markovien classique (modèle de mots connectés avec contraintes de durée) mis au point dans le projet Esprit Multiworks (1989-1993). L'effort que nous mettons dans la robustesse ne porte pas sur l'amélioration de ce moteur mais sur l'ensemble des modules qui prétraitent ou postraitent les données. Ce qui contribue à la robustesse, ce sont tous les composants placés autour de ce moteur, à la fois "en haut" (contraintes linguistiques), "à gauche" (traitement des entrées multimodales), "en bas" (vérification phonétique et prosodique), et "à droite" (rejet et postraitement).

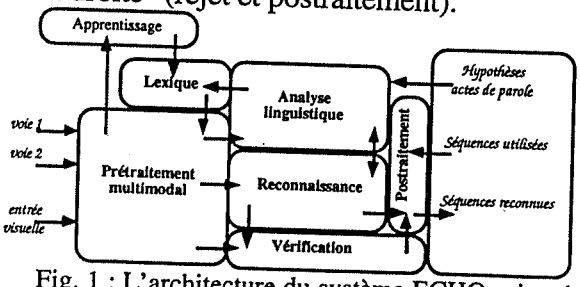


Fig. 1 : L'architecture du système ECHO orienté vers la robustesse : dans cette architecture l'accent est mis sur l'ensemble des modules qui entourent le moteur de reconnaissance proprement dit (voir détails dans le texte).

Cette architecture est l'aboutisement d'une réflexion menée à propos de la "robustesse" de la communication verbale chez l'humain. Entre locuteurs humains, on parle avec une intention, celle de communiquer ou celle d'agir [Sperber 86]. Parler c'est émettre une séquence sonore, selon un code convenu, dans un environnement acoustique donné et dans un monde dans lequel ces événements prennent une signification. Ecouter, c'est pour le destinataire collecter ces morceaux de code sonore en fonction d'un but ou d'une action qu'il présuppose — que ce but soit partagé ou non avec son (ou ses) interlocuteur(s). Pour collecter ces morceaux de code il doit non seulement reconnaître les éléments de ce code mais aussi les ré-organiser pour les comprendre vis-à-vis de la situation vécue. En situation adverse, l'auditeur humain utilise encore plus les contraintes offertes par le code linguistique pour réduire les hypothèses et les ambiguïtés, mais aussi utilise les informations provenant de la situation : événements concomitants, événements redondants, etc. Généralement l'auditeur sait dans quel environnement il est plongé et en connait les limitations. Il sait aussi la tâche qu'on lui donne ou qu'il se donne : il est motivé par un but qu'il cherche à atteindre. Reconnaître n'est donc qu'un des maillons de la chaîne des processus qu'il met en jeu pour agir. Ce maillon est intégré et entrelacé avec les autres processus avec lequel il participe. Reconnaître, signifie donc chez l'humain, identifier des éléments déjà appris mais aussi posséder des mécanismes qui permettent de généraliser à des éléments nouveaux les processus de reconnaissance acquis pour des éléments anciens. C'est aussi reconnaître qu'il y a des éléments omis ou des éléments redondants, s'adapter à l'environnement, au locuteur, tenir compte de l'évolution de la situation, etc. Bref, pour l'auditeur humain, la reconnaissance n'est l'un que des très nombreux processus qui l'amènent à réagir de manière adéquate à un acte de parole de son interlocuteur.

Replacées dans le cadre de la communication homme-machine, ces remarques conduisent à considérer :

a) que la machine devrait être pourvue d'intentions, afin de guider sa recherche d'éléments pertinents dans l'énoncé, de les repérer, de lever les ellipses et de désambiguïser certains énoncés,

b) que la machine devrait être pourvue d'une conscience, lui permettant d'évaluer ses propres résultats, capacité qui concourt à lui permettre de s'améliorer, de s'adapter, d'apprendre,

c) que la machine devrait vivre dans le monde et en percevoir les changements, pour mesurer les effets de ses actions et en

retirer des jugements qui lui permettent de s'améliorer.

Evidemment de telles capacités sont illusoires pour la machine, ou du moins elles ne peuvent conduire qu'à des artefacts si on tente de les atteindre. Cependant, traduites dans un autre langage, les remarques ci-dessus peuvent aider à structurer un système robuste :

a) les deux processus de reconnaissance et de compréhension (ou dialogue) doivent se répondre de manière réflexive et complémentaire dans le décours du dialogue,

b) il faut doter la machine de capteurs pour lui permettre de "d'évaluer" le résultat de ses actions dans le monde pour en avoir des connaissances d'arrière-plan, ou de posséder des représentations explicites de la tâche pour appréhender totalement la situation correspondant à l'énonciation,

c) il faut donner à la machine des possibilités de s'évaluer, ou lui fournir en continu des évaluations de ses réponses. A partir de là, elle pourra par exemple relancer un apprentissage ou affiner ses modèles.

L'architecture que nous proposons (Fig. 2) tient compte de ces réflexions :

a) le système dispose d'entrées multimodales (pour capter l'environnement et utiliser la redondance des informations dans des conditions difficiles),

b) le système de reconnaissance possède un niveau de rétroaction avec un système de dialogue et de compréhension,

c) la cohérence avec le code linguistique est vérifiée à tous les niveaux : phonétique, prosodique, sémantique,

d) l'adaptativté est résolue par réflexivité avec les systèmes environnants.

Dans les sections suivantes nous décrivons brièvement quelques uns de ces modules.

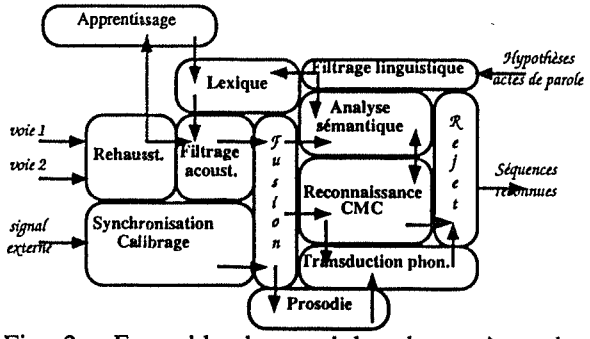


Fig. 2 : Ensemble des modules du système de reconnaissance robuste ECHO :

1. Reconnaissance par modèles de Markov (CMC)
2. Filtrage acoustique (élimination des sons non vocaux)
3. Filtrage linguistique (élimination des séquences invalides)
4. Lexique (de mots et modèles acoustiques)
5. Rehaussement de la parole
6. Rejet
7. Analyse sémantique
8. Synchronisation, calibrage de sources
9. Fusion des informations issues de différents capteurs
10. Transduction phonétique (repérage de certains phonèmes)
11. Prosodie (contrôle des marques prosodiques)

3. LA ROBUSTESSE AMONT

3.1. Le filtrage acoustique

Les sources de perturbation acoustique sont nombreuses et variées. Les bruits peuvent être familiers ou inattendus mais ils sont rarement évitables. Il vaut mieux plutôt les intégrer au processus de reconnaissance que d'essayer de les éliminer. Dans le cas où ils sont familiers on peut les intégrer de la manière suivante :

a) "apprendre" et étalonner ces bruits dans diverses conditions d'émission en variant la distance du micro et l'intensité de la source B (on suppose dans un premier temps que le rayonnement du bruit est équivalent à celui d'une source ponctuelle),

b) puis "apprendre" l'espace sonore relativement à la position de la source utile U placée à une distance d de la source B. On montre pour cela qu'il suffit d'étalonner le système en un nombre fini de positions choisies dans le champ sonore. En effet si le champ d'intensité acoustique est en 1/r, les lignes de champ sont des hyperboles de foyers B et U. Il suffit donc de placer le micro en un nombre fini de positions entre B et U pour échantillonner complètement l'espace,

c) on constitue ainsi un dictionnaire de sons et de sources étalonnées qui permet ensuite en fonctionnement nominal (1) le repérage des sources dans l'espace, (2) l'évaluation du rapport signal/bruit, (3) la constitution de gabarits de masquage pour l'apprentissage des modèles markoviens.

Les bruits inattendus sont généralement intenses et brefs puisqu'ils correspondent à des incidents. Un détecteur de bruits intenses est souhaitable pour inhiber la chaîne de reconnaissance.

3.2. Le rehaussement du signal

Nous avons montré dans des travaux précédents [NguyenThi 94] que deux cas sont à considérer pour rehausser le signal noyé dans le bruit :
a) le rapport signal/bruit ≥ 5 dB,
b) le rapport signal/bruit < 5dB.

Dans le premier cas une méthode de soustraction du bruit peut suffire, mais dans le second cas une méthode par séparation de sources est préférable. Dans ce cas le modèle de Jutten-Hérault donne des performances acceptables (24 dB de gain si le modèle additif peut être employé et 12 dB dans le cas d'un modèle convolutif — le choix de l'un ou l'autre de ces modèles dépend de la nature acoustique de l'environnement, et de la distance des sources).

Le module de rehaussement du signal est le suivant (on suppose ici que les sources X et Y sont convoluées au cours de leur propagation acoustique avant d'atteindre les deux microphones placés à une distance d l'un de l'autre) :

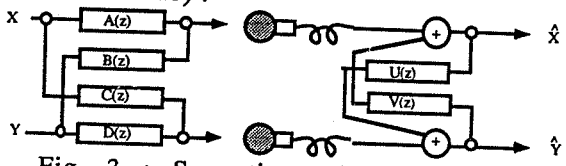


Fig. 3 : Synoptique de la méthode de rehaussement de la parole par séparation de sources. Les sources X et Y sont supposées se mélanger dans les deux microphones selon des lois convolutives. Les signaux captés sont "filtrés" par un réseau de neurones adaptatif à une seule couche U(z) et V(z) que l'on optimise sur le critère de minimisation des cumulants croisés d'ordre 4. La sortie donne deux signaux approchant X et Y.

L'adaptation du système est continue ce qui permet de traiter des sources évolutives dans le temps. Le gain obtenu est de l'ordre de 12 à 24 dB [Nguyen Thi, 94]. Les taux de reconnaissance augmentent dans ces conditions d'environ 20% (de 60% à 80% en milieu fortement bruité).

3.3. La fusion multimodale

Il est bien connu que des sources d'information additonnelles améliorent la robustesse de la reconnaissance dans le bruit (mouvement des lèvres, gestes, mouvements du visage, etc.) [Robert-Ribes 95]. Diverses tentatives ont été faites pour fusionner au micux ces informations dans un système markovien en vue de la reconnaissance (projet AMIBE [Montacié 95] notamment). Kabré a montré [Kabré 96] qu'il est préférable de les fusionner en amont du système de reconnaissance plutôt que dans le moteur lui-même. On gagne alors 10 à 15% en conditions difficiles.

3.4. La parole spontanée

Le traitement de la parole spontanée (pauses intempestives, hésitations, reprises, etc.) ne peut être traitée valablement qu'au niveau linguistique. Antoine [Antoine, 94] propose diverses solutions au niveau micro-sémantique mais ces solutions n'ont pu encore être intégrées au système ECHO en raison de leur complexité. Ce travail est en cours.

4. LA ROBUSTESSE AVAL

4.1. La cohérence phonétique

Pour des modèles de mots, la cohérence phonétique des résultats obtenus par des méthodes markoviennes n'est pas toujours assurée. En effet parmi les listes d'hypothèses retenues il peut y en avoir qui présentent des aberrations évidentes au regard de certains traits ou indices acoustiques. Il est alors assez facile de les éliminer. Pour cela la méthode est la suivante :

a) on considère les r meilleures séquences hypothèses obtenues par le système,

b) on génère les séquences phonétiques correspondantes (il y a en général p>r),

c) puis on génère les séquences acoustiques à partir de modèles de phones, et,

d) on compare par programmation dynamique (algorithme DTW pondéré) la chaîne ainsi produite et la chaîne originale.

Cet algorithme [Lichene, 92] permet de ré-ordonner les r meilleures hypothèses de départ selon un ordre de cohérence décroissant et non plus selon le critère des probabilité décroissantes puisque l'on sait que ce critète n'est pas significatif vis-à-vis d'un rejet efficace. Ici robustesse égale cohérence.

4.2. La cohérence prosodique

De nombreuses études ont été faites pour tenter d'utiliser la prosodie en reconnaissance de la parole [Nasri 89]. Nos recherches récentes dans le système MICRO

[Caillaud 94] nous ont montré la possibilité d'utiliser la prosodie dans une phase de vérification des frontières de mots et de syntagmes. De la liste des r séquences précédentes on déclasse celles dont les frontières sont mal placées.

4.3. Le rejet : un processus réflexif

Le rejet résulte d'un double processus : le rejet amont (par le système lui-même, comme aux §4.2 et §4.3) et le rejet aval (par le système utilisateur, par exemple le système de dialogue). L'idée est de profiter du rejet aval pour donner au système de reconnaissance, des points de repère pour s'améliorer.

Le rejet amont opère un filtrage des hypothèses,

- sur les r meilleures probabilités,
- par ré-ordonnancement sur la cohérence phonétique et prosodique.

Le rejet aval est effectué par le système qui utilise ces résultats de reconnaissance. L'analyse des fréquences des séquences ou des mots non utilisés permet de relancer l'apprentissage de ces mots ou de ces séquences. Il est en effet facile de faire une statistique des mots et des séquences de la manière suivante :

supposons qu'à l'instant t(n) la liste ordonnée des séquences soit,

s(1,n) au rang r(1), s(2,n) au rang r(2), etc. et que le système de dialogue ait utilisé la séquence s(i,n) fournie au rang r(i), alors on cumule P(j) <- P(j) + r(j) pour tout j tel que r(j)<r(i). Dès que P(j) dépasse un certain seuil cela prouve que la séquence s(j) est trop souvent ignorée bien que placée à un rang privilégié. On en déduit que cette séquence est probablement trop favorisée, soit parce qu'elle est trop peu discriminante, soit parce qu'elle ne reflète pas l'énoncé à reconnaître : dans les deux cas il y a lieu de refaire l'apprentissage en ligne avant de continuer le dialogue en cours (ou de ré-étalonner le niveau acoustique).

5. CONCLUSION

Le système ECHO auquel nous avons abouti présente une plus grande robustesse, ce mot étant pris selon plusieurs sens :

a) tout d'abord dans le sens d'une meilleure cohérence des résultats vis-à-vis du contenu linguistique et phonétique et des attentes dialogiques,

b) ensuite vis-à-vis des conditions variées de l'environnement acoustique,

c) enfin par rapport à l'adaptatibilité du système qui est capable de relancer des apprentissages en fonction des réultats produits et donc de maintenir un taux de performance constant.

La robustesse ainsi envisagée ne conduit pas à des performances de pointe accrues mais à une meilleure utilisabilité du système c'est-à-dire à des performances plus régulières quelles que soient les conditions d'utilisation du système.

6. REFERENCES

[Antoine 94] J.Y. Antoine, Coopération syntaxe-sémantique pour la compréhension automatique de la parole spontanée. Thèse SIP, INPGrenoble, 1994.

[Caillaud 94] B. Caillaud, J.Y. Antoine, J. Caelen, G. Caelen-Haumont, MICRO : Un système multi-agents pour la compréhension de la parole. Congrès RFIA, AFCET, Nancy, 1994.

[Delemar 95] O. Delemar, Un modèle mixte de reconnaissance de la parole, thèse INPG, Grenoble, 1996 (soutenance prochaine).

[Gong 95] Y. Gong, Speech recognition in noisy environments : a survey. Speech Com., 16, 1995, p. 261-291.

[Haton 95] J.P. Haton, Modèles neuronaux et hybrides en reconnaissance de la parole : état des recherches, , in Fondements et Perspectives en TAP, H. Méloni éd., 1995, p. 139-154.

[Kabré 96] H. Kabré, A probalistic model for multi-sensor data fusion, Speech Com. (accepté).

[Lichene 92] A. Lichene, Vérification d'hypothèses phonétique. Rapport de DEA-SIP, Grenoble, 1992.

[Montacié 95] Cl. Montacié et al., Applications Mutimodales pour Interfaces et Bornes Evoluées (AMIBE), in Fondements et Perspectives en TAP, H. Méloni éd., 1995, p. 155-164.

[Nasri 89] M.K. Nasri, G. Caelen-Haumont et J. Caelen, 1989, "Using Prosodic Rules in Continuous Speech Recognition", Proc. of ICASSP-IEEE, Glasgow, Vol. 1, pp. 671-674.

[NguyenThi 94] L. NguyenThi, Rehaussement de la parole par séparation de sources, Thèse INPG, Grenoble, 1994

[Robert-Ribes 95] J. Robert-Ribes, Modèles d'intégration audiovisuelle de signaux linguistiques : de la perception humaine à la reconnaissance des voyelles. Thèse INPG-SIP, Grenoble, 1995.

[Rossi 95] M. Rossi, Quelles connaissances utiliser en reconnaissance de la parole ? , in Fondements et Perspectives en TAP, H. Méloni éd., 1995, p. 201-210.

[Sperber 86] D. Sperber and D. Wilson, La pertinence. Editons de Minuit, Paris, 1986.

TECHNIQUES DE COMPENSATION POUR LA RECONNAISSANCE DE LA PAROLE BRUITÉE

Driss MATROUF, Jean-Luc GAUVAIN

LIMSI-CNRS, BP 133, F-91403 Orsay, France
Tél: 69 85 80 67 - Fax: 69 85 80 80 - e-mail: {driss,gauvain}@limsi.fr

ABSTRACT

The performance of speech recognizers degrade substantially when there is a mismatch between the training and testing conditions. The goal of noise compensation is to minimize the effects of the mismatch, so as to bring the recognition performance as close as possible to the obtained under matched conditions. Different approaches to achieve robustness have been studied. These approaches may be split into two groups. The first class of approaches make use of a channel model for additive and convolutional noises. Techniques in this category include spectral substraction[3], cepstral mean substraction[10], noise masking, the CDCN algorithm[9], speech and noise decomposition[2], and parallel model combination[4][7][8]. The second class of approaches makes no assumptions about the underlying noise, and uses some optimality criterion (generally Maximum Likelihood (ML)) to find a transformation that is applied either to the signal or to the acoustic models used by the recognizer. Techniques in this category include ML linear regression[11][6] and ML stochastic matching[5]. In this paper we discuss the properties of above mentioned techniques. The results of our analyses have led to the implementation of noise compensation in the LIMSI large vocabulary CSR system. Evaluation results on the Nov95 ARPA NAB CSR test data are given both with and without noise compensation.

INTRODUCTION

La présence de bruit engendre une dégradation significative des performances des systèmes de reconnaissance de la parole, en particulier lorsque les conditions d'apprentissage et de test sont différentes. Les techniques de compensation ont pour but de réduire les effets de cette différence et d'obtenir un taux d'erreur comparable à celui obtenu lorsque les conditions d'apprentissage et de test sont identiques. Les techniques utilisées pour appréhender le problème du bruit peuvent se répartir en deux classes.

La première classe repose sur un modèle du canal de transmission avec deux types de bruits: additif et convolutif, i.e. le signal observé est $y(t) = (s(t) + n(t)) * h(t)$, où $n(t)$ et $h(t)$ désignent respectivement les bruits additif et convolutif. Au sein de cette classe, on peut distinguer deux approches différentes. La première consiste à estimer $s(t)$ à partir du signal bruité $y(t)$ et de statistiques sur $h(t)$ et $n(t)$. Parmi les techniques suivant cette approche, on peut citer la soustraction spectrale[3], la soustraction du cepstre moyen[10], le masquage de bruit et l'algorithme CDCN (Codeword-Dependent Cepstral Normalization)[9]. La deuxième approche consiste à adapter les modèles retenus pour le signal propre $s(t)$ afin d'obtenir des modèles représentatifs du signal bruité $y(t)$. On peut citer la décomposition de modèles[2], et la combinaison parallèle des modèles[4][7][8].

La deuxième classe ne suppose aucun modèle *à priori* du bruit, mais utilise un critère d'optimalité généralement basé sur le maximum de vraissemblance pour estimer une transformation à appliquer sur le signal ou sur les modèles acoustiques du système de reconnaissance. Dans cette classe, on peut citer l'adaptation par régression linéaire (MLLR: Maximum Likelihood Linear Regression)[6]et la compensation stochastique[5].

L'objectif de cette étude est de faire une analyse permettant de choisir les techniques les plus pertinentes à incorporer dans le système de reconnaissance du LIMSI. Cette analyse est basée soit sur des résultats expérimentaux, soit sur une étude théorique qui met en évidence les limites des techniques non retenues. Seuls les résultats des tests validant les techniques retenues sont présentés. Des tests comparatifs avec les autres techniques ne sont pas présentés car d'une part les évaluations des différentes techniques ont été généralement réalisées sur des corpus différents et d'autre part, la place impartie ne permet pas d'exposer l'ensemble des résultats avec les conditions expérimentales correspondantes.

TECHNIQUES BASÉES SUR UN MODÈLE *À PRIORI*

Dans cette section nous analysons les techniques fondées sur un modèle *à priori*. Les modèles de bruits généralement utilisés sont: $y(t) = s(t) + n(t)$, $y(t) = s(t) * h(t)$ ou $y(t) = (s(t) + n(t)) * h(t)$, avec $n(t)$ et $h(t)$ désignent respectivement les bruits additif et convolutif et $y(t)$ le signal observé.

Soustraction spectrale

Cette technique consiste à effectuer une décomposition spectrale de chaque trame du signal bruité. Puis chaque canal du spectre est atténué selon que le niveau mesuré localement dans le canal dépasse plus ou moins l'estimation du bruit. La densité spectrale du bruit est estimée dans les périodes de silence[3]. L'utilisation d'un seuil pour éviter les valeurs négatives causées par la variance du bruit et d'un facteur de surestimation pour augmenter le rapport signal sur bruit[12] introduisent des formes spectrales complétement inconnues du système de reconnaissance de la parole qui se base principalement sur une classification des formes spectrales à court terme du signal de parole. En plus, la non prise en compte de l'effet du bruit sur les variances des modèles limite considérablement l'utilisation de cette technique dans le cadre de la reconnaissance automatique de la parole.

Soustraction du cepstre moyen

Cette technique consiste à soustraire le cepstre moyen à long terme. En général on soustrait le cepstre moyen sur la phrase. Cette idée est couramment utilisée pour éliminer les distortions liées au changement du canal d'enregistrement[10]. Après soustraction cepstrale on aboutit à des paramètres cepstraux indépendants du matériel d'enregistrement utilisé. Il est clair que la moyenne estimée dépend de la proportion du silence (bruit) dans la phrase. Cette dépendance est indésirable. Pour résoudre ce problème, un mélange de deux gaussiennes a été utilisé: une pour les trames de parole et une autre pour les trames du silence (bruit)[13]. Cette classification se fait en utilisant l'algorithme EM avec comme paramètre de classification l'énergie[1]. Ainsi la soustraction du cepstre moyen se fait comme suit:
$\hat{y}^c(t) = y^c(t) - [\gamma\bar{y}_p^c + (1-\gamma)\bar{y}_b^c]$, $y^c(t)$ est la représentation cepstrale d'une trame à l'instant t. γ désignant la probabilité que $y^c(t)$ soit de la parole. $\bar{y}_p^c$ et $\bar{y}_b^c$ désignent respectivement la moyenne des vecteurs cepstraux de parole et du silence.

Principe du masquage et la décomposition parole/bruit

L'idée du masquage du bruit[2], appelé aussi l'approximation du MAX peut se résumer par la formule:
$Y_t^l = log(X_t + N_t) \simeq max[X_t^l, N_t^l]$, Y_t^l représente le niveau du logarithme de l'énergie dans un canal d'un banc de filtres de la parole bruitée à l'instant t, X_t^l et N_t^l représentent respectivement le niveau du logarithme de l'énergie de la parole et du bruit dans un canal donné. C'est cette idée qui est à l'origine de toutes les méthodes de compensation actuellement utilisées dans les systèmes de reconnaissance. En effet, ce même principe a été étendu pour donner lieu à la technique de la décomposition de la parole et du bruit (SND: Speech and Noise Decomposition)[2]. Il s'agit d'une méthode optimale pour reconnaître la parole et le bruit simultanément. L'approximation du MAX faite par Varga[2] donne une bonne estimation de la densité de probabilité de la parole bruitée. Cependant, elle suppose que le bruit agit indépendamment d'un canal à l'autre dans le domaine log-spectral. En plus cette technique impose l'utilisation des matrices de covariance pleines pour modéliser la corrélation entre les composantes.

Combinaison parallèle des modèles

La combinaison parallèle des modèles (PMC) [7][8] a été introduite pour surmonter les limitations de la SND. Elle est cependant directement inspirée de cette dernière. Elle procède à toutes les opérations inverses pour revenir du domaine cepstral au domaine spectral (changement des densités de probabilités suivant l'opération) et inversement.

L'approximation log-normale[7][8] est souvent utilisée pour revenir du domaine spectral au domaine cepstral. Elle suppose que la somme de deux variables aléatoires log-normalement distribuées est elle-même log-normalement distribuée. Il est important de noter que l'approximation log-normale devient inacceptable dans le cas de grandes variances, c'est la raison pour laquelle il faut répartir l'espace acoustique en plusieurs sous-espaces de petites variances.

La différence principale entre la SND et la PMC est le type d'approximation: la PMC utilise l'approximation log-normale ou procède à une intégration numérique, par contre la SND utilise l'approximation du MAX définie précédemment. Une autre différence est que la SND opère dans le domaine log-spectral alors que la PMC opère dans le domaine cepstral. Il est important de noter que la SND augmente considérablement le nombre de paramètres dans le système. La difficulté des calculs et le nombre d'approximations dans le cadre de la PMC augmentent rapidement avec la complexité des formules utilisées pour les paramètres cepstraux différentiels, ce qui représente un grand inconvénient de l'approche.

Compensation du bruit convolutif dans le cadre de la PMC

L'idée retenue est la soustraction du cepstre moyen. Après avoir appliqué la PMC, on obtient des modèles correspondant à la parole bruitée. Pour obtenir des modèles pour la parole bruitée avec soustraction du cepstre moyen, il suffit de soustraire le cepstre moyen de toute la parole bruitée à toutes les moyennes des gaussiennes composant les modèles. Mais it à priori on ne dispose pas des données bruitées pour pouvoir calculer cette moyenne. Pour résoudre ce problème, on utilise une somme de gaussiennes représentant toute la parole propre; on fait la composition de celle-ci avec le modèle de bruit dont on dispose pour obtenir un modèle composé de plusieurs gaussiennes représentant toute la parole bruitée. Grâce à ce modèle on peut estimer d'une manière satisfaisante le cepstre moyen de la parole bruitée qu'on peut ensuite soustraire à toutes les moyennes des gaussiennes. Il est important de noter que l'utilisation d'un mélange de gaussiennes pour représenter toute la parole propre est nécessaire car une seule gaussienne aurait une très grande variance et l'approximation log-normale de la PMC ne sera plus valable dans ce cas. Des tests effectués au LIMSI montrent que cette technique donne des résultats tout-à-fait satisfaisants.

Utilisation de la composition directement sur les données

Pour résoudre certains problèmes liés à la PMC, on utilise une nouvelle technique développée au LIMSI[14] utilisant la composition directement sur les données. Elle consiste à stocker pour chaque gaussienne d'un état donné, d'un modèle donné, les vecteurs cepstraux qui ont permis son estimation. La composition consiste dans ce cas à faire la somme trame à trame dans le milieu spectral des trames stockées pour la gaussienne en question et des trames représentant le bruit. Les trames représentant le bruit proviennent généralement d'une analyse à court terme du signal recueilli dans les périodes de silence au moment du test. On applique les opérations nécessaires pour revenir au domaine cepstral. On obtient ainsi les vecteurs cepstraux qui permettent de réestimer la moyenne et la variance de la gaussienne en question. Comme pour la PMC, on suppose que le bruit n'affecte pas la distribution des trames entre les gaussiennes d'un état donné. Il est clair qu'il s'agit d'une supposition dont la validité dépend du type de bruit et de son niveau. Pour surmonter ce problème, il est possible d'adopter le même raisonnement au niveau de l'état et non au niveau de la gaussienne. Ceci implique une reclassification des trames (en utilisant l'algorithme EM) dans chaque état. Il est également possible d'utiliser l'ancienne classification comme initialisation de l'algorithme EM. Ainsi seul le calcul nécessaire pour ajuster les modèles est effectué. Cette technique est similaire à Data Driven PMC (DDPMC) utilisée par Gales et Young[8], la différence principale est que la technique DDPMC regénère aléatoirement les vecteurs associés à chaque état au lieu de les stocker au moment de l'apprentissage. Ceci est nécessairement coûteux en temps de calcul si on veut générer un nombre suffisant de vecteurs pour avoir une validité statistique.

La composition directement sur les données permet aussi de réaliser la soustraction du cepstre moyen. Pour faire ceci correctement, il faut estimer h ou n à partir du bruit $h * n$[14]. Le bruit peut être estimé d'une manière itérative partant des trames n_0 de silence dans les données de test. Ces trames de silence peuvent être utilisées pour calculer le cepstre moyen de la parole bruitée, qui est ensuite soustrait au cepstre moyen des données d'adaptation pour obtenir une première estimation de $\tilde{h}$. Le filtre $\tilde{h}^{-1}$ est alors appliqué aux données d'adaptation pour obtenir une meilleure estimation de n. En pratique 5 itérations suffisent pour avoir une bonne estimation de h.

TECHNIQUES BASÉES SUR UN CRITÈRE D'OPTIMALITÉ

Ces techniques ne supposent en général aucun modèle *à priori*. Elle sont basées sur un critère d'optimalité, généralement le maximum de vraisemblance. Parmi les travaux les plus importants dans ce cadre, on peut citer la régression linéaire (MLLR)[6][11].

Le système est adapté à un nouvel environnement par des transformations linéaires des paramètres. Les transformations sont estimées en alignant les données d'adaptation avec les états des modèles. Il est supposé qu'on ne dispose que d'une petite quantité de données dans le nouvel environnement, ce qui rend impossible l'adaptation individuelle de chaque paramètre du modèle (les données ne couvrent pas d'une manière suffisante la totalité des modèles). Pour cela on procède à une classification des paramètres où chaque classe subira la même transformation.

La moyenne d'une densité gaussienne est adaptée en utilisant une transformation linéaire W. W est une matrice $(n \times (n+1))$ contenant éventuellement une colonne supplémentaire pour modéliser *l'offset*. La moyenne adaptée est donnée par: $\mu_{ad} = W\hat{\mu}$. La valeur choisie pour W est celle qui maximise la vraisemblance que les données d'adaptation soient générées par les modèles adaptés. La technique utilisée pour ceci est l'algorithme EM, c'est-à-dire, la détermination de la fonction auxiliaire puis sa maximisation. D'une

Tab. 1 - *Taux d'erreur (sur les mots) avec et sans technique de compensation pour les deux types de données (C0: propres, P0: bruitées).*

PMC compens.	*MLLR adapt.*	*Taux d'erreur (%)*	
		P0	*C0*
non	non	>50.0	10.4
oui	non	20.5	10.4
oui	oui	17.5	9.1

manière similaire, on peut adapter les variances.

RÉSULTATS ET CONCLUSIONS

Les résultats de ces analyses nous ont amenés à incorporer certaines techniques de compensation du bruit dans le système de reconnaissance de parole continue du LIMSI[14] qui utilise un vocabulaire de 65.000 mots. Nous présentons ici les résultats obtenus sur les données de test ARPA NAB CSR de novembre 1995 qui ont été enregistrées dans un environement bruité (47 à 61dBA). Ces données de test comprennent 300 phrases prononcées par 20 locuteurs (15 phrases par locuteur,$\cong$ 20 mots/phrase). Chaque énoncé a été enregistré simultanément avec deux microphones: un microphone Sennheiser HMD-410 (rapport S/B moyen égal à 30dB), et un microphone inconnu du système (rapport S/B de 7 à 18dB).

Le même système de reconnaissance, entraîné uniquement sur des données propres enregistrées avec un microphone Sennheiser HMD-410, a été utilisé pour décoder les deux ensembles de données (C0: Sennheiser microphone, P0: autres microphones). Les résultats en terme de taux d'erreur sur les mots sont donnés dans le tableau 1. On peut observer que l'utilisation des techniques de compensation améliore d'une manière significative les performances sur les données bruitées. La technique PMC compense le bruit additif lié à l'environnement et le bruit convolutif lié au microphone. Le grand avantage de cette technique est qu'elle améliore de manière très significative les performances dans le cas de présence de bruit additif ou convolutif (> 50%/20.5% d'erreurs), tout en gardant les mêmes performances dans le cas d'absence de bruit (10.4%/10.4% d'erreurs). L'adaptation MLLR compense les différences résiduelles entre les données d'apprentissage et celles du test correspondant d'une part au bruit non representé dans le modèle du canal de transmission, et d'autre part, à la variabilité interlocuteur. Une seule matrice de regression (49 × 48) a été utilisée pour transformer les moyennes des modèles. D'après la dernière ligne du tableau 1, on constate que la technique MLLR apporte des améliorations dans tous les cas. Ces résultats, obtenus avec un système de reconnaissance de parole continue, indépendant du locuteur, avec un vocabulaire de 65.000 mots (vocabulaire de test illimité), montrent la pertinence des techniques de compensation retenues.

RÉFÉRENCES

[1] D. V. Compernolle, "Noise adaptation in a hidden Markov model speech recognition system," *Computer Speech & Language*, pp. 151-167, 1989.

[2] A.P. Varga, R.K. Moore, " Hidden Markov model decomposition of speech and noise," *ICASSP-90*, pp. 845-848.

[3] S.F Boll, " Suppression of Acoustic Noise in Speech Using Spectral Substraction," *IEEE*, pp. 113-120, 1979.

[4] F. Martin, K. Shikano, Y. Minami, "Recognition of Noisy Speech by Composition of Hidden Markov Models," *EuroSpeech'93.*

[5] A. Sankar, C.-H. Lee, "Robust Speech Recognition based on Stochastic Matching," *ICASSP-95*, pp. 121-124.

[6] C.J. Leggetter, P.C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech & Language*, pp. 171-185, 1995.

[7] M.J.F. Gales, S.J. Young, "Robust speech recognition in additive and convolutional noise using parallel model combination," *Computer Speech & Language*, pp. 289-307, 1995.

[8] M.J.F. Gales, S.J. Young, "A fast and flexible implementation of parallel model combination," *ICASSP-95*, pp. 133-136.

[9] A. Acero, R.M. Stern, "Enviromental Robustness in Automatic Speech Recognition," *IEEE Acoustics, Speech & Signal Processing*, pp. 849-852. April 1990.

[10] B. Atal, "Effectiveness of Linear Prediction Characteristics of Speech Wave for Automatic Speaker Identification and Verification," *Journal of the Acoustic Society of America*, **55**, pp. 1304-1312. June 1974.

[11] O. Siohan, Y. Gong, J-P. Haton, "Noise adaptation using linear regression for continuous noisy speech recognition" *EuroSpeech'95*, pp. 465-468.

[12] O. Cappé, "Techniques de réduction de bruit pour la restauration d'enregistrements musicaux" *thèse TELECOM Paris 93 E 019*, sept 93.

[13] X. Huang, A. Acero, F. Alleva, M-Y. Hwang, L. Jiang and M. Mahajan "Microsoft windows highly intelligent speech recognizer: whisper" *ICASSP-95*, pp. 93-96.

[14] J.L. Gauvain, L. Lamel, G. Adda, D. Matrouf, "Developments in Continuous Speech Dictation using the 1995 ARPA NAB News Task," *ICASSP-96.*

UN NOUVEL ALGORITHME DE RECHERCHE DANS LES RESEAUX DE SEGMENTATION MULTI-NIVEAUX

Jean-Luc HUSSON, Yves LAPRIE

CRIN-CNRS & INRIA Lorraine, BP 239 - 54506 Vandœvre-Lès-Nancy

Tél.: 83 59 20 91 - Fax: 83 41 30 79 - e-mail: husson@loria.fr,

ABSTRACT

This paper describes a new segmentation system using a multi-level representation, called a dendrogram. We address the issues of estimating the confidence of one path, and finding the N most reliable paths in the segmentation lattice. Our approach rests on automatically trained criteria and on an efficient strategy to prune the search space.

1. INTRODUCTION

Dans les systèmes analytiques de D.A.P., la réussite de la phase d'identification phonétique est fortement conditionnée par la qualité de la phase de segmentation qui précède, ce qui justifie les recherches consacrées à la conception d'algorithmes fiables et robustes de segmentation automatique. Parmi tous les algorithmes proposés, nombreux sont ceux qui reposent sur la recherche de discontinuités dans le signal ou dans son spectre au cours du temps. Ces algorithmes présentent cependant les inconvénients suivants :

- Ils ne fournissent en général qu'une solution unique de segmentation.
- Ils se heurtent à la difficulté de calculer localement le seuil qui minimise globalement les effets de sur ou de sous-segmentation du signal.
- Ils n'assurent pas l'homogénéité des unités obtenues. Aussi devient-il impossible d'utiliser en aval de la segmentation un système de décodage n'utilisant qu'un unique type d'unités phonétiques.

Pour pallier ces inconvénients, des algorithmes générant des réseaux de segmentation multi-niveaux, notamment les dendrogrammes (Glass, 1988), ont été proposés. Ces réseaux de segmentation sont calculés à partir de la représentation spectrale d'un signal de parole. La (figure 1) montre une représentation stylisée d'un tel dendrogramme.

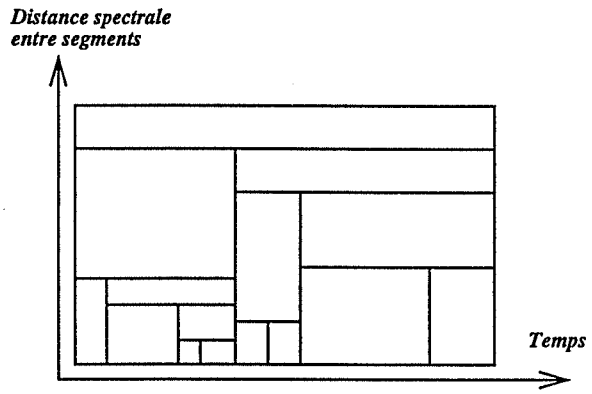


Figure 1 : Exemple de dendrogramme.

Les avantages de cette structure sont multiples :

- Ils représentent potentiellement toutes les décompositions possibles du signal de parole (de la plus grossière à la plus fine), dans une structure hiérarchique uniforme, couvrant ainsi toutes les valeurs significatives de seuils possibles. Une étude préalable (Hajislam, 1994) a montré que la bonne segmentation (celle de l'expert) était toujours présente dans le dendrogramme.
- Les segments sont automatiquement classés hiérarchiquement en fonction de leur homogénéité acoustique.
- La segmentation correspondant au dendrogramme est indépendante de tout niveau de représentation segmentale car tous les niveaux sont potentiellement représentés dans la structure, des événements subphonémiques aux mots.

La contrepartie de ces avantages est la complexité du dendrogramme obtenu. Le grand nombre de chemins qu'il contient (de l'ordre de 2 milliards de segmentations pour 2 secondes de parole) rend impossible l'examen exhaustif de toutes les solutions par le module d'identification. Une sélection automatique des

meilleurs chemins dans cette structure est alors indispensable, mais en dépit de quelques tentatives (Hübener, 1993), ce problème n'a pas reçu de solution satisfaisante à ce jour.

Nous proposons ici un algorithme permettant de résoudre ce problème et de trouver les solutions de segmentation d'un dendrogramme les plus vraisemblables. Nous nous intéresserons successivement aux problèmes suivants :

- Evaluation de la vraisemblance d'un chemin.
- Recherche automatique des N meilleurs chemins dans un dendrogramme, au sens du maximum de vraisemblance.

2. VRAISEMBLANCE D'UN CHEMIN

Nous choisissons de calculer la confiance accordée à un chemin C du dendrogramme comme le produit des confiances des segments qui le composent. En pratique, nous considérons le logarithme de cette confiance :

$$\log[p(C)] = \sum_{S_i \in C} \log[p(S_i)]$$

La confiance $p(S)$ d'un segment est obtenue en combinant deux probabilités correspondant chacune à un critère d'évaluation de sa vraisemblance phonétique :

- Son homogénéité acoustique
- La probabilité de sa durée

2.1 L'homogénéité acoustique

Intuitivement, un segment est d'autant moins probable qu'il est peu homogène spectralement. Nous considérerons donc la probabilité d'appartenance d'un segment à la segmentation finale en fonction de son homogénéité spectrale. Celle-ci est mesurée par la différence des spectres moyens (vecteurs représentants) des fils du segment considéré.

Le vecteur représentant d'un segment est un vecteur de 12 coefficients cepstraux (MFCC), correspondant à la moyenne des vecteurs MFCC calculés toutes les 8 ms sur le segment, pondérés par un coefficient décroissant avec leur éloignement au centre de celui-ci. L'apprentissage de la probabilité a été réalisé par la modélisation d'une variable aléatoire gaussienne réelle à 12 dimensions (nombre de coefficients MFCC) en utilisant la segmentation manuelle et les vecteurs différence précédemment décrits, calculés pour tous les segments des dendrogrammes de notre corpus d'apprentissage.

2.2 La probabilité de durée du segment

On cherche à estimer la vraisemblance de la durée d_S d'un segment S sous la forme d'une probabilité $p_{\text{durée}}(S)$. Notre première tentative a consisté à calculer la probabilité d'observer un segment d'une telle durée sur la base d'un apprentissage de la distribution des durées *tous sons confondus* du corpus d'apprentissage. Ce critère grossier s'est révélé peu sélectif en raison de la grande disparité des durées observées. Afin de rendre le critère de durée plus pertinent, nous avons choisi de procéder à une classification phonétique préalable du segment et de retenir la valeur maximale des probabilités d'observer cette durée pour les phonèmes h de la classe phonétique prédite $c_{max}(S)$ du segment :

$$p_{\text{durée}}(S) = \max_{h \in c_{max}(S)} p\langle d_S | h \rangle$$

Pour ce faire, nous avons dû mettre en oeuvre une procédure de ***classification automatique*** des segments (cf. 2.2.1) et ***modéliser pour chaque phonème du français la distribution des durées*** observées sur le corpus d'apprentissage (cf. 2.2.2).

2.2.1 Classification automatique

La procédure de classification mise en œuvre a pour objectif de classer un segment S dans l'une des 6 classes phonétiques de la partition suivante :

θ =*{voyelles_orales, voyelles_nasales, fricatives, occlusives, semi-voyelles, sonantes}*

On attribue à chaque segment S l'hypothèse la plus vraisemblable de classe $c_{max}(S)$:

$$c_{max}(S) = \underset{c \in \theta}{\operatorname{argmax}}[p_{class}\langle c|S\rangle]$$

avec

$$\begin{cases} p_{class}\langle c|S\rangle = \dfrac{p\langle S|c\rangle \cdot p(c)}{p(S)} \\ p(S) = \displaystyle\sum_c p\langle S|c\rangle \cdot p(c) \end{cases}$$

Les 6 classes ont été modélisées par des variables aléatoires gaussiennes réelles à 12 dimensions (nombre de coefficients MFCC du vecteur représentant des sons). Les paramètres de chaque gaussienne sont calculés sur la base

d'un apprentissage réalisé à partir d'une base globale d'environ 7000 sons. Les probabilités d'observation des classes $p(c)$ sont issues d'études statistiques présentées dans (Tubach, 1985).

2.2.2 *Modélisation de la durée des phonèmes*

La probabilité $p\langle d_S|h\rangle$ d'observer la durée d_S pour une réalisation d'un phonème h a été modélisée par une fonction gamma (Burshtein, 1995) pour tous les phonèmes du français sur la base d'un corpus d'apprentissage d'environ 7000 segments étiquetés manuellement.

La modélisation ainsi obtenue est à ce jour indépendante du contexte et de la position du phonème dans le groupe rythmique. Cependant, une procédure de normalisation de la durée par le débit d'élocution est en cours de mise au point. Cette normalisation devrait augmenter la précision de la modélisation des durées et donc le pouvoir sélectif de ce critère.

3. ALGORITHME DE RECHERCHE

Pour choisir les meilleurs chemins, nous avons développé un algorithme de programmation dynamique qui explore le dendrogramme. Cet algorithme évalue la vraisemblance de tous les chemins possibles de la manière décrite précédemment et construit de proche en proche les N meilleurs chemins au sens du maximum de vraisemblance. Cette stratégie de construction itérative des meilleures solutions rend le temps de traitement proportionnel à la durée du signal à traiter bien que le nombre de chemins croisse de manière exponentielle avec celle-ci.

Le dendrogramme contient cependant un grand nombre de chemins peu vraisemblables mais pris néanmoins en compte dans le parcours. Aussi, deux contraintes complémentaires ont-elles été ajoutées à l'algorithme pour restreindre l'espace de recherche aux hypothèses vraisemblables. La première contrainte est une ***contrainte locale de voisement*** (cf. 3.1) qui élimine certains segments phonétiquement irréalistes. La seconde contrainte est une ***contrainte globale de durée*** qui élimine les chemins qui présentent un nombre de segments trop grand ou trop petit (cf. 3.2).

3.1 La contrainte locale de voisement

Il est clair que les meilleurs chemins doivent respecter les frontières de voisement car un son ne peut être à la fois voisé et non voisé. Parmi les différentes stratégies testées, nous avons retenu celle qui consiste à décomposer le dendrogramme D de l'énoncé en une suite contiguë de sous-dendrogrammes de D qui respectent *au mieux* les frontières de voisement préalablement déterminées par l'algorithme de pitch (Martin, 1982). Cette technique permet d'éliminer dans le dendrogramme global les segments peu vraisemblables vis-à-vis du critère de voisement, en particulier les très longs segments, ce qui diminue la complexité du problème. De plus, la décomposition du dendrogramme global en une suite de dendrogrammes plus courts renforce l'efficacité de la seconde contrainte (contrainte globale de durée) décrite dans le paragraphe suivant (cf 3.2).

Cependant, cela nous a contraint à modifier l'algorithme de sélection des meilleurs chemins globaux afin de prendre en compte les résultats obtenus localement sur les sous-dendrogrammes. Cette partie du travail ne sera pas décrite ici.

3.2 La contrainte globale de durée

Cette contrainte indique, en fonction de la durée de l'énoncé, le nombre de segments attendus sous la forme d'un intervalle de confiance. Plutôt que de l'appliquer dès le début pour élaguer l'espace complet de recherche, cette contrainte est générée et prise en compte dynamiquement pendant le parcours, sur les portions de l'énoncé exclusivement voisées ou non voisées. Nous décrivons ci-dessous la procédure de construction de l'intervalle de confiance.

Pour mettre en œuvre cette contrainte, nous avons dû au préalable modéliser les distributions de durées de toutes les zones de 1 à 15 segments (sans changement d'état de voisement) par une gaussienne. Ce choix de 15 segments au maximum a été fixé par apprentissage. Il permet de couvrir 99.7 % des observations avec des gaussiennes significatives (avec un nombre d'échantillons suffisant pour obtenir une bonne représentation de l'histogramme des effectifs observés par la gaussienne correspondante). Le faible nombre de données disponibles pour les zones de plus de 15 segments ne permet pas de construire les gaussiennes correspondantes.

Nous pouvons alors calculer la probabilité $p\langle k|d\rangle$ d'observer une zone de k segments connaissant sa durée d grâce à la relation suivante:

$$\begin{cases} p\langle k|d\rangle = \dfrac{p\langle d|k\rangle \cdot p(k)}{p(d)} \\ p(d) = \displaystyle\sum_k p\langle d|k\rangle \cdot p(k) \end{cases}$$

$p\langle d|k\rangle$ est fournie directement par la $k^{ième}$ gaussienne et $p(k)$ est la part de zones de k segments sur le nombre total de zones, obtenue par simple comptage dans le corpus d'apprentissage.

Pour chaque zone de durée d, un intervalle I_d est construit sur le principe de base suivant :

$$\left(k_{max} = \underset{k \in [1,15]}{\arg max} \; [p\langle k|d\rangle] \right) \in I_d$$

... et est itérativement étendu de manière à satisfaire simultanément les deux conditions suivantes :

$$\begin{cases} \displaystyle\sum_{k \in I_d} p\langle k|d\rangle > 1 - \varepsilon \\ \text{Largeur}(I_d) \text{ minimale} \end{cases}$$

Notons que cet intervalle existe toujours. L'étude de l'évolution de la taille moyenne des intervalles ainsi construits, en fonction de l'erreur ε de prédiction a priori tolérée, a permis de fixer ce paramètre à moins de 1 %. La largeur moyenne des intervalles obtenus est de 4,5 segments. Cette largeur est très inférieure à l'écart-type moyen du nombre de segments des chemins présents dans les sous-dendrogrammes, ce qui confère une grande efficacité à cette contrainte.

3.3 Effet des contraintes

Les deux contraintes apportent un double avantage :

- Elles suppriment les segments et les chemins peu probables, contribuant à l'efficacité de la fonction de sélection.
- Elles diminuent la complexité de la structure à gérer, réduisant le temps nécessaire à son traitement informatique.

De plus, les tests réalisés montrent que l'adjonction de ces deux contraintes n'ajoute pas de risque d'erreur.

4. RESULTATS ET DISCUSSION

Nous nous sommes intéressés au problème de la recherche des N-meilleures solutions de segmentation dans les structures multi-niveaux (dendrogrammes). Contrairement à d'autres algorithmes qui utilisent des règles heuristiques, notre approche repose sur un apprentissage systématique ou sur une adaptation automatique des paramètres de notre système.

Les tests préliminaires de comparaison manuelle des segmentations obtenues automatiquement et des étiquetages des experts valident la méthode proposée et démontrent l'efficacité de l'algorithme de recherche des meilleures solutions de segmentation. Les bons résultats obtenus nous engagent à procéder à une évaluation automatique de ce système sur de vastes corpus du français, grâce à un algorithme d'alignement automatique de segmentations.

Cet algorithme peut être facilement adapté au traitement d'autres structures hiérarchiques multi-niveaux (Witkin, 1983). Notre système pourra être complété par la simple adjonction de critères et de contraintes supplémentaires. La prise en compte du contexte devrait notamment permettre d'affiner la pertinence des critères de vraisemblance déjà mis en œuvre.

5. BIBLIOGRAPHIE

Glass J. R., V. W. Zue (1988) "Multi-Level Acoustic Segmentation of Continuous Speech", *Proc. ICASSP-88, pp. 215-218*

Hajislam R. (1994) "Décodage acoustico-phonétique et robustesse en reconnaissance automatique de la parole", *Thèse de doctorat de l'université Henri Poincaré-Nancy I*

Hübener K., A. Hauenstein (1993) "Controlling search in segmentation lattices of speech signals", *Proc. EUROSPEECH '93, V. 3, pp.1763-1766*

Tubach J.-P., L.-J. Boe (1985) "*Un corpus de transcriptions phonétiques : constitution et exploitation statistiques*", Rapport ENST 85D001

Burshtein D. (1995) "Robust parametric modeling of durations in hidden Markov models", *Proc. ICASSP-95, pp. 548-551*

Martin Ph. (1982) "Comparison of pitch detection by cestrum and spectral comb analysis", *Proc. ICASSP-82, pp. 180-183*

Witkin A. P. (1983) "Scale-space filtering", *IJCAI-83, pp.1019-1022*

UTILISATION DE MODÈLES DE MARKOV POUR L'ÉTIQUETAGE AUTOMATIQUE ET LA RECONNAISSANCE DE BREF80

Dominique Fohr, Jean-François Mari, Jean-Paul Haton

CRIN-CNRS & INRIA Lorraine, Bâtiment LORIA, BP 239, 54506 Vandoeuvre-lès-Nancy

Tél.: 83.59.20.00 - Fax: 83.41.30.79 - e-mail: {fohr,jfmari,jph}@loria.fr

ABSTRACT

This paper presents recent work on continuous speech labelling. A method based on second order HMM is presented and assessed on TIMIT. Compared to a manual labelling, results show that 93% of borders are set within an interval of 30ms. Various problems of transcription from orthographic form to phones are discussed. In section 4., we give phonetic recognition rates on BREF80 after the labelling of this corpus.

1. INTRODUCTION

Au fur et à mesure que les performances des systèmes de reconnaissance de la parole augmentent, il est nécessaire de les tester sur des ensembles de plus en plus volumineux. Dans le cadre de la reconnaissance du discours continu en français, le corpus BREF80 (Lamel, 1991) constitué de 5330 phrases lues par 80 locuteurs est un outil puissant d'évaluation et de comparaison de systèmes de reconnaissance de la parole.

Dans cet article, nous décrivons notre méthode pour utiliser ce corpus afin de tester notre système de décodage acoustico-phonétique basé sur des modèles de Markov cachés d'ordre 2 (HMM2 pour reprendre l'abréviation anglaise). Comme cette mise en oeuvre s'appuie sur un étiquetage préalable de ce corpus en phonèmes, nous décrirons notre méthode de segmentation automatique qui utilise, elle aussi, des HMM2.

2. DESCRIPTION DE BREF80 ET ANALYSE ACOUSTIQUE

BREF80 est un corpus de parole lue. 80 locuteurs ont lu des phrases issues du journal Le Monde sans préciser la ponctuation. En tout 5330 phrases ont été enregistrées à 16 kHz sur 2 CDROMs.

Pour paramètrer le signal, nous utilisons 12 coefficients MFCC calculés toutes les 8ms auxquels nous ajoutons des coefficients de régression du premier ordre et du deuxième ordre.

3. DESCRIPTION DES MODÈLE

Dans un modèle du second ordre (HMM2) la suite d'états cachés est supposée être une suite de Markov du second ordre, c'est-à-dire que la probabilité de transition entre états dépend des états dans lesquels le modèle se trouvait aux temps t-1 et t-2.

3.1. Notations

On appellera:

- λ le modèle de Markov d'ordre 2,
- S_t la variable aléatoire au temps t dont les réalisations sont les états du modèle,
- b_i la densité associée à l'état i,
- O_t l'observation au temps t (de dimension D),
- $P(O/\lambda)$ la vraisemblance de la suite d'observations $O_1,O_2...O_T$ avec le modèle λ,
- $\aleph(\mu, \Sigma)$ une loi normale de dimension D de moyenne μ et de matrice de covariance Σ,
- on notera M^t la transposée de la matrice M.

3.2. Passage de l'ordre 1 à l'ordre 2

Classiquement, les probabilités de transition entre états d'un HMM1 sont :

$$\text{Prob}(S_t = k \mid_{S_{t-1} = j,\, S_{t-2} = i,\, S_{t-3} = \ldots}) = \text{Prob}(S_t = k \mid_{S_{t-1} = j}) = a_{jk}$$

Dans un HMM2 elles deviennent :

$$\text{Prob}(S_t = k \mid_{S_{t-1} = j,\, S_{t-2} = i,\, S_{t-3} = \ldots}) = \text{Prob}(S_t = k \mid_{S_{t-1} = j,\, S_{t-2} = i}) = a_{ijk}$$

Les densités associées aux états restent multinormales

$$b_i(O_t) = \sum_m c_{im}\, \aleph(\mu_m, \Sigma_m, O_t) \qquad \sum_m c_{im} = 1$$

$$\aleph(\mu, \Sigma, x) = \frac{1}{\sqrt{(2\pi)^D \det \Sigma}} e^{-\frac{1}{2}(x-\mu)^t \Sigma^{-1}(x-\mu)}$$

Les probabilités "Forward-Backward" s'obtiennent simplement en ajoutant un indice indiquant l'état d'où on vient :

$$\alpha_{t+1}(j,k) = \sum_{i=1}^{S} \alpha_t(i,j) a_{ijk} b_k(O_{t+1}), \quad 2 \le t \le T-1$$

$$\beta_t(i,j) = \sum_{k=1}^{S} \beta_{t+1}(j,k) a_{ijk} b_k(O_{t+1}), \quad 2 \le t \le T-1$$

Les compteurs de transitions deviennent:

$$\eta_t(i,j,k) = \frac{\alpha_t(j,k) a_{ijk} b_k(O_{t+1}) \beta_{t+1}(j,k)}{P(O/\lambda)}, \quad 1 \le t \le T-1$$

A partir de ces définitions, la réestimation, suivant le maximum de vraisemblance, des paramètres des modèles est classique (Mari, 1994).

4. ETIQUETAGE DE BREF80

Notre but est de réaliser l'apprentissage d'un HMM2 pour chaque unité phonétique en utilisant un corpus de parole dont on n'a que la transcription orthographique. L'apprentissage se fait en trois étapes:

- étiquetage phonétique manuel d'une petite partie du corpus,
- apprentissage des premiers modèles de Markov à l'aide du seul petit corpus précédent,
- étiquetage automatique de l'ensemble du corpus grâce aux premiers modèles,
- apprentissage des modèles plus précis à l'aide de l'ensemble du corpus.

Nous allons détailler les phases de l'étiquetage.

4.1. Étiquetage phonétique manuel

Pour réaliser l'étiquetage manuel d'un petit corpus, nous avons utilisé le logiciel SNORRI (Fohr, 1989) développé au CRIN. Ce logiciel permet, entre autre, d'écouter des morceaux de signal, de visualiser le signal temporel et le spectrogramme et de poser des étiquettes phonétiques.

4.2. Étiquetage automatique à partir d'une chaîne de phonèmes

Pour étiqueter automatiquement des fichiers de parole, nous utilisons un modèle de Markov appris à l'aide du petit corpus étiqueté à la main. La segmentation est immédiate si on

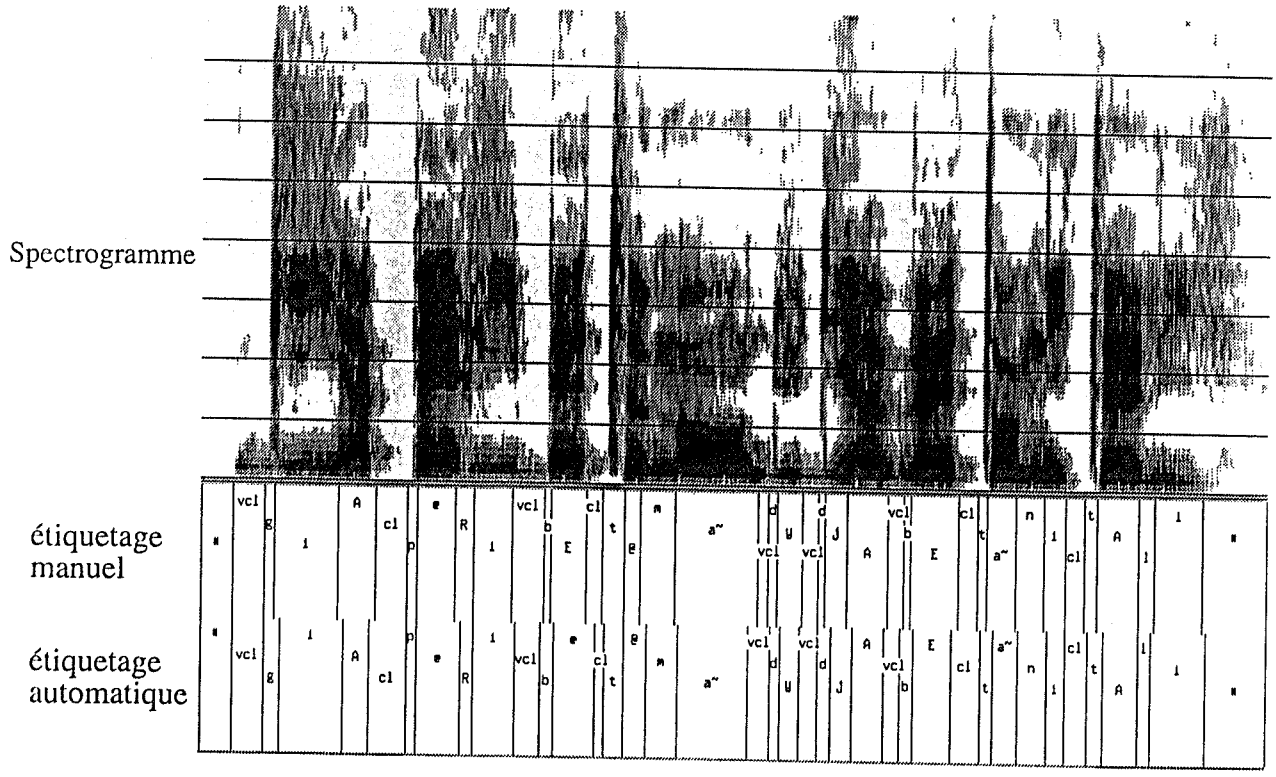


Figure 1. : Opération d'étiquetage de la phrase "Guy a péri bêtement du diabète en Italie"

possède la transcription en phonèmes de la phrase. Il suffit d'utiliser l'algorithme de Viterbi comme le montre la figure 1.

4.3. Performances

Pour évaluer les performances de cette approche, il est nécessaire de posséder un corpus étiqueté qui va servir de référence. Nous avons choisi la base de données américaine TIMIT (Garofolo, 1993), un corpus entièrement étiqueté manuellement. A l'aide de modèles de phonèmes (Mari, 1996), nous avons étiqueté les 1344 phrases de test. Les résultats sont donnés table 1.

Table 1: répartition des frontières en fonction de l'écart entre la segmentation manuelle et la segmentation automatique

% de frontières dont l'écart est <= à 20 ms	86.1%
% de frontières dont l'écart est <= à 30ms	93.2%
% de frontières dont l'écart est <= à 40ms	96.2%
% de frontières dont l'écart est <= à 50ms	97.7%
% de frontières dont l'écart est <= à 67ms	99.0%

Dans plus de 93% des cas, la frontière est placée à moins de 30 ms de la frontière placée par l'expert du MIT, ce qui nous paraît un bon résultat compte tenu de la difficulté de la tâche et du fait qu'un nombre important de frontières ont été placées de façon arbitraire.

Ce résultat montre que, si on est capable de produire la bonne séquence de phonèmes, notre procédure permet de placer automatiquement et de façon fiable les frontières phonétiques.

4.4. Etiquetage automatique à partir d'une suite de mots

Dans le cas de BREF80, les phrases proviennent de la lecture d'articles du journal Le Monde et nous ne possédons que le texte écrit de l'article. Il est nécessaire d'effectuer au préalable une transcription de graphèmes en phonèmes. A l'aide du dictionnaire BDLEX (Pérennou, 1992), qui contient la transcription phonétique, la finale phonologique et des règles de liaison, nous pouvons générer un ensemble de transcriptions possibles.

4.5. Transcription graphèmes-phonèmes: problèmes rencontrés

Le problème de la transcription graphème-phonème pour l'étiquetage est sensiblement différent de celui posé pour la synthèse de parole. En effet, pour la synthèse, il suffit de trouver UNE prononciation acceptable. Pour l'étiquetage, il faut aligner LA prononciation que le locuteur a utilisée. Par exemple, «1900» peut se prononcer «mille neuf cents» ou «dix neuf cents». Il faut donc générer toutes les prononciations possibles d'une phrase, les aligner et celle qui donne le meilleur score d'alignement est utilisée pour réaliser l'étiquetage.

Voici une liste, non exhaustive, des problèmes rencontrés lors de l'étiquetage de BREF80:

- les pauses entre les mots et les «e muet» doivent être systématiquement prévus ;
- les liaisons doivent être prévues ;
- pour les mots d'origine étrangère et les noms propres, plusieurs prononciations, même étranges, doivent être imaginées. Par exemple, l'ex-président Américain Bush a été prononcé: /buʃ/ (ce qui semble naturel) /byʃ/ et même /bøʃ/ !
- les acronymes peuvent s'épeler ou se prononcer comme un seul mot: par exemple notre laboratoire le CRIN ;
- pour les chiffres, il faut tenir compte du fait que certaines finales sont optionnelles, 500 peut se prononcer «sɛ̃k sã» ou sɛ̃ sã»;
- la virgule doit être traitée différemment suivant qu'elle représente un signe de ponctuation ou une virgule décimale: dans le dernier cas elle se prononce par le mot virgule ou point (pour les locuteurs anglophones);
- certaines initiales peuvent, suivant le contexte se prononcer de manière très différente : par exemple V peut être prononcé «cinquième» dans V république, «vé» dans V. Dupont et «Victor» dans V. Hugo.

Toutes ces difficultés montrent à quel point il est difficile d'envisager une procédure purement automatique pour l'étiquetage.

Nous proposons de définir plusieurs critères pour trouver les étiquetages erronés:

- rejeter les phrases qui s'alignent mal lors de l'apprentissage,
- calculer la probabilité *a posteriori* d'un phonème connaissant ses limites. Si une suite de phonèmes a une probabilité faible, cette phrase est aussi rejetée.

Toutes les phrases ainsi rejetées sont transcrites à l'écoute par un expert et sont ensuite réalignées automatiquement.

5. RÉSULTATS DE RECONNAISSANCE SUR BREF80

Dans un premier temps, nous avons étiqueté manuellement un total de 650 phrases prononcées par 60 locuteurs. A l'aide de ces phrases, 35 modèles de phonèmes à trois états comportant chacun 5 densités normales ont été appris.

Ces premiers modèles ont servi à étiqueter les deux CD de BREF80. Parallèlement à cette activité, nous avons testé un premier système de reconnaissance comportant ces 35 modèles et une grammaire bigramme de phonèmes. En l'absence de partition apprentissage/test, nous avons utilisé 20 locuteurs pour le test n'appartenant pas au corpus d'apprentissage. Les résultats au niveau phonétique apparaissent table 2.

Table 2: résultats de reconnaissance sur BREF80 à l'aide de 35 modèles

test sur 200 phrases 10 locuteurs (20000 phonèmes)	%
corrects	74.0 %
omissions	11.6 %
insertions	1.5 %
précision (accuracy)	72.5 %

Nous pouvons comparer ces résultats avec ceux obtenus par L. Lamel et J.L Gauvain dans (Lamel, 1993) qui annoncent 78.7 % de reconnaissance. Toutefois, il faut noter qu'ils utilisent 428 modèles dépendants du contexte alors que nous n'utilisons que 35 modèles indépendants du contexte.

Nous avons testé sur TIMIT le même système et observé un résultat de 70.3%, ce qui est moindre que pour le français. Un écart analogue avait été observé par L. Lamel et J.L Gauvain.

6. CONCLUSIONS ET PERSPECTIVES

Pour réaliser un système de dictée automatique, il est nécessaire d'augmenter les performances du système acoustique en utilisant des modèles plus précis comme des modèles dépendants du contexte ou plus grand nombre de densités par états. Ceci implique l'estimation d'un plus grand nombre de paramètres et donc l'utilisation de plus gros corpus qui ne pourront plus être étiquetés manuellement.

C'est pour cette raison que nous avons développé un outil permettant l'étiquetage automatique d'un grand corpus de parole continue. Dans 93 % des cas, la frontière phonétique placée par le système est à moins de 30 ms de celle placée par un expert humain lorsque la chaîne phonétique est connue. Ce résultat peut être considéré comme suffisant.

7. BIBLIOGRAPHIE

Lamel L.F., Gauvain J.L., Eskénazi M. (1991) "BREF, a Large Vocabulary Spoken Corpus for French", *Eurospeech*, 505-508

Mari J.F., Fohr D., Haton J.P. (1994) "Modèles stochastiques d'ordre 1 et 2", *JEP*, 199-202

Fohr D., Laprie Y. (1989) "Snorri: an Interactive Tool for Speech Analysis", *Eurospeech*, 1:613-616

Garofolo J.S., Lamel L.F., Fisher W.M., Fiscus J.G., Pallett D.S., Dahlgren N.L. (1993) "The DARPA TIMIT Acoutic-Phonetic Continuous Speech Corpus CDROM", NTIS order number PB91100354

Mari J.F., Fohr D., Junqua J.C. (1996) "A Second-Order HMM for High Performance Word and Phoneme-Based Continuous Speech Recognition", *IEEE ICASSP*

Pérennou G., Cotto D., De Calmes M., Ferrané I., Pecatte J.M. (1992) "Le projet BDLEX de base de données lexicale du français écrit et parlé", *Séminaire Lexique*, IRIT-UPS, Toulouse

Lamel L.F., Gauvain J.L. (1993) "High Performance Speaker-Independent Phone Recognition Using CDHMM", *Eurospeech*, 121-124.

UTILISATIONS D'UNE SEGMENTATION A PRIORI DU SIGNAL DE PAROLE DANS UN SYSTEME DE RECONNAISSANCE PAR MODELES DE MARKOV CACHES

Thierry MOUDENC, Jean MONNE

France Télécom - CNET - LAA/TSS/RCP - 2 Av. Pierre Marzin -22307 LANNION Cedex

Tél: 96 05 37 16- Fax: 96 05 35 30 - e-mail: moudenc@lannion.cnet.fr

ABSTRACT

This paper presents the speech recognition experiments conducted on using the stationarity changes of the speech signal. Two approaches have been explored. The first one is based on the introduction of a new parameter in the acoustic vectors used in the Markovian modeling, while the second one makes use of a N-best solutions segmental post-processing.

Each of these approaches enables an error rate reduction. However, the best improvement results from their combination. The speaker independent recognition experiments conducted on isolated words telephone databases lead to an error rate reduction, with respect to the basic HMM model, from 8% up to 27 %, depending on the database.

1. INTRODUCTION

Nous nous intéressons ici à l'utilisation de l'information provenant des ruptures de stationnarité dans le signal de parole. La figure 1 donne une illustration de ces ruptures. Le signal de parole représenté en haut de la figure correspond à une prononciation du mot "Perros". Les flèches indiquent les instants de rupture dans la stationnarité du signal. La détection des instants de rupture de stationnarité est réalisée par la méthode de divergence Forward-Backward (Obrecht, 1988).

Dans une approche markovienne classique, les vecteurs d'observations sont extraits du signal de parole sur des fenêtres de largeur constante illustrées sur la figure par les cadres en traits discontinus. Ces vecteurs sont ensuite alignés sur les modèles des phonèmes qui composent le mot reconnu. Comme on peut le voir, les zones stationnaires, délimitées par 2 ruptures de stationnarité consécutives, peuvent s'étaler sur des durées beaucoup plus larges que celles des fenêtres d'analyse.

Une approche dans l'utilisation de ces ruptures de stationnarité consiste à effectuer une analyse LPC sur chacune des zones stationnaires pour en extraire un vecteur acoustique. Ces vecteurs sont ensuite utilisés avec un modèle de Markov (Farhat, 1994).

Dans le travail que nous présentons, l'utilisation des instants de rupture de stationnarité est réalisée à travers deux méthodes. La première consiste en l'ajout d'un nouveau paramètre aux vecteurs acoustiques. Ce paramètre représente la position temporelle

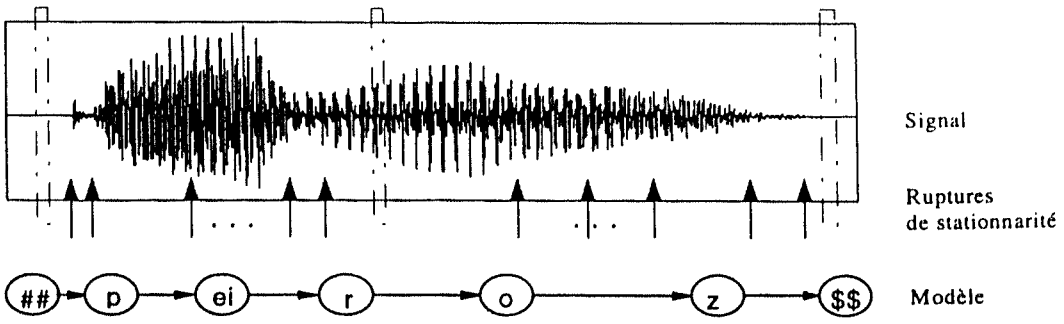


Figure 1 : Exemple de ruptures de stationnarité dans le signal de parole

relative du centre de la trame courante par rapport aux ruptures de stationnarité précédente et suivante dans le signal de parole. La seconde (Moudenc, 1995) consiste en l'estimation de modèles du nombre de ruptures de stationnarité par segment phonétique. Ces modèles sont ensuite utilisés dans le cadre d'un post-traitement des N meilleures solutions.

2. SYSTEME DE BASE

2.1 Le système de reconnaissance Phil90

Le système de reconnaissance que nous utilisons est le système Phil90 (Jouvet, 1994), dédié à la reconnaissance de parole de qualité téléphonique en mode indépendant du locuteur. Il est basé sur une modélisation markovienne du signal de parole. Chaque mot est représenté par un réseau phonétique traduisant les différentes prononciations possibles et chaque phonème est représenté par un modèle de Markov contextuel (Jouvet, 1991).

Les vecteurs acoustiques utilisés sont constitués de 27 coefficients. Il s'agit des 8 premiers coefficients cepstraux, de l'énergie ainsi que de leurs dérivées temporelles premières et secondes. Les coefficients cepstraux sont extraits, durant la phase d'analyse acoustique, toutes les 16 ms, sur des trames de signal d'une largeur de 32 ms.

2.2 Bases de données

Les expérimentations ont été menées sur 3 bases de données de parole de qualité téléphonique enregistrées chacune par environ 800 locuteurs : la base des *Chiffres,* composée des chiffres de 0 à 9, la base des *Nombres*, composée des nombres de 00 à 99, et la base du *Trégor*, composée de 36 mots ou expressions de la langue française. Chaque base est découpée en 2 parties de taille similaire : un ensemble d'apprentissage et un ensemble de test. Le tableau 1 fournit le détail du nombre d'enregistrements (mots ou expressions) contenus dans chacun des ensembles de test ainsi que les taux d'erreur correspondants en utilisant les vecteurs acoustiques décrits précédemment. L'objectif étant ici la reconnaissance sans rejet, chaque élément du vocabulaire (chiffre, nombre ou expression) doit être reconnu à part entière et l'on ne mesure que le taux de substitution.

Tableau 1 : Nombre d'enregistrements et taux d'erreur pour chaque ensemble de test.

	Chiffres	*Nombres*	*Trégor*
Taille corpus de test	3622	7288	12842
Taux d'erreurs	1,13 %	3,57 %	0,83 %
Intervalles de confiance	0,83:1,52	3,17:4,02	0,68:1,0

3. INTRODUCTION D'UN PARAMETRE ACOUSTIQUE COMPLEMENTAIRE

La segmentation du signal de parole par la méthode de divergence Forward-Backward (Obrecht, 1988) fournit les instants de rupture dans la stationnarité du signal.

Pour chaque trame de parole est calculée la position temporelle relative du centre de la trame courante par rapport aux ruptures de stationnarité précédente et suivante. Cette valeur est introduite dans les vecteurs acoustiques comme paramètre supplémentaire. Un apprentissage classique est alors effectué par l'algorithme de Viterbi pour estimer les paramètres du modèle de Markov utilisé. On notera la simplicité de mise en oeuvre puisque, en dehors de la détermination des instants de rupture de stationnarité, il s'agit de traiter un vecteur acoustique composé de 28 coefficients au lieu de 27.

Sur les ensembles de test des différents corpus de parole utilisés, les réductions de taux d'erreur de reconnaissance obtenues par rapport au système de base sont reportées dans le tableau 2 :

Tableau 2 : Réductions des taux d'erreur après introduction du paramètre de segmentation a priori dans les vecteurs acoustiques.

	Chiffres	*Nombres*	*Trégor*
Taux d'erreurs	0,83 %	3,57 %	0,72 %
Réductions des taux d'erreur	**26 %**	**2 %**	**13 %**

4. POST-TRAITEMENT SEGMENTAL DES N-MEILLEURES SOLUTIONS

Une précédente étude (Moudenc, 1995) avait porté sur l'utilisation d'un post-traitement des N meilleures solutions et permis une réduction sensible des taux d'erreur. L'approche est basée sur l'estimation de modèles statistiques du nombre de ruptures de stationnarité par segment phonétique. Nous en rappelons le principe ci-dessous.

4.1 Formalisme utilisé

Pour un mot à reconnaître, le principe consiste à calculer, pour chacune des N meilleures solutions markoviennes, un score de post-traitement segmental Sc_{PTS_i} pour un jeu i de coefficients segmentaux. Ce score est donné par l'équation :

$$Sc_{PTS_i}(\text{Align}) = \text{Log}\left(\frac{Pr_i(\text{Align est correct})}{Pr_i(\text{Align est incorrect})}\right)$$

où Align représente l'alignement markovien de la solution et $Pr_i(.)$ représente la probabilité d'émission de la solution sachant une modélisation du jeu i de coefficients segmentaux. Ainsi, Pr_i(Align est correct) (resp. incorrect) représente la probabilité que la solution qui correspond à l'alignement Align est la bonne solution (resp. une solution erronée).

Ce score est ensuite combiné linéairement avec le score markovien, ce qui fournit le score final de chaque solution :

$$Sc(\text{Align}) = \alpha_0 Sc_{HMM}(\text{Align}) + \sum_{i=1}^{NbPTS} \alpha_i Sc_{PTS_i}(\text{Align})$$

où NbPTS représente le nombre de post-traitements segmentaux mis en oeuvre et Sc_{HMM} représente le score Markovien obtenu par la solution. Les coefficients de pondération α_i sont quant à eux estimés par l'algorithme de Powell (NRC, 1990) sur l'ensemble d'apprentissage.

4.2 Modélisation segmentale

Des densités de probabilité discrètes du nombre de ruptures de stationnarité par segment phonétique sont estimées sur les ensembles d'apprentissage. Cette modélisation est effectuée pour les segments phonétiques corrects -issus des alignements des solutions correctes- et pour les segments phonétiques "incorrects", issus des alignements incorrects.

Les segments phonétiques sont modélisés d'une part en fonction du contexte d'apparition (i.e. des contextes gauches et droits) et d'autre part indépendamment du contexte. Un lissage, dont les paramètres sont appris au maximum de vraisemblance, est ensuite effectué entre les scores des modélisations contextuelles et hors contextes.

De plus, afin de considérer les ruptures de stationnarité situées près des frontières des segments, la modélisation est paramétrée par un pas de tolérance sur les frontières markoviennes. Ce pas permet ainsi d'élargir (pas > 0) ou de rétrécir (pas < 0) les segments (Moudenc, 95).

4.3 Résultats de base du post-traitement

Les réductions de taux d'erreur à l'issue du post-traitement, relatives aux taux obtenus par le système de base, sont reportées dans le tableau 3. La deuxième ligne correspond à la prise en compte du pas de tolérance qui fournit la meilleure réduction d'erreur sur l'ensemble d'apprentissage. La dernière ligne correspond à la prise en compte simultanée de 5 pas de tolérance choisis sur l'ensemble d'apprentissage combinés linéairement et dont les paramètres de combinaison sont estimés par l'algorithme de Powell.

Tableau 3 : Réductions des taux d'erreur après application du post-traitement segmental.

		Chiffres	*Nombres*	*Trégor*
Taux d'erreurs obtenus par le système de base		1,13 %	3,57 %	0,83 %
Réductions des taux d'erreur	1 seul pas de tolérance	**15 %**	**6 %**	**9 %**
	Combinaison des scores de 5 pas de tolérance	**22 %**	**12 %**	**5 %**

Tableau 4 : Réductions des taux d'erreur obtenues pour les différentes méthodes de prise en compte de la segmentation a priori.

		Chiffres	*Nombres*	*Trégor*
Taux d'erreurs obtenus par le système de base		1,13 %	3,57 %	0,83 %
Réductions correspondantes	Post-traitement avec 5 pas de tolérance	**22 %**	**12 %**	**5 %**
	Introduction du paramètre de segmentation a priori dans les vecteurs acoustiques	**26 %**	**2%**	**13 %**
	Paramètre de segmentation a priori dans les vecteurs acoustiques + post-traitement	**27 %**	**8 %**	**11 %**

5. COMBINAISON DES DEUX METHODES

La combinaison des deux méthodes consiste, dans un premier temps, à introduire le paramètre issu de la segmentation a priori, décrit en partie 3, dans les vecteurs acoustiques puis à entrainer le modèle de Markov utilisé pour la reconnaissance et à déterminer les N meilleures solutions markoviennes. On applique ensuite le post-traitement segmental que nous venons de décrire sur ces N meilleures solutions.

Les réductions de taux d'erreurs obtenues à l'issue du traitement complet sont reportées dans le tableau 4, où sont aussi rappelés les résultats précédents.

CONCLUSION

Les résultats obtenus montrent clairement que l'utilisation des instants de rupture de stationnarité du signal de parole permet une amélioration des performances de reconnaissance. De plus, il est intéressant de remarquer que la combinaison des deux méthodes permet de tirer le meilleur profit de l'information segmentale.

REMERCIEMENTS

Les auteurs tiennent à exprimer leur gratitude envers R. André-Obrecht qui a fourni les programmes de segmentation a priori et à D. Jouvet pour ses précieux conseils. Nous remercions le Conseil Régional de Bretagne pour son soutien financier.

REFERENCES

Farhat A., Parlangeau N., Obrecht R. (1994), Deux systèmes d'étiquetage phonétique basés sur une pré-segmentation du signal de parole, *Proc. JEP 94*, Trégastel, France, pp. 169-173.

Jouvet D., Monné J., Bartkova K., Mokbel C. (1994), Structure des modèles de Markov et coefficients acoustiques, *Séminaire Reconnaissance Automatique de la Parole,* Nancy, France.

Jouvet D., Bartkova K., Monné J. (1991), On the modelisation of allophones in an HMM based speech recognition system, *Proc. EUROSPEECH 91*, Gênes, Italie, *pp. 923-926.*

Moudenc T., Jouvet D., Monné J. (1995), On Using A-Priori Segmentation of the Speech Signal in an N-Best Solutions Post-Processing, *Proc. ICASSP 95*, Detroit, USA, *pp. 580-583.*

NRC (1990), *Numerical Recipes in C*, Cambridge University Press.

Obrecht R. (1988), A New Statistical Approach for the Automatic Segmentation of Continuous Speech Signal, *IEEE Trans. on ASSP, Vol. 36, pp. 29-40.*

VALIDATION DE TRAITS PHONETIQUES PAR UN SYSTEME DE RECONNAISSANCE DE L'ARABE STANDARD

Sid-Ahmed SELOUANI*, Jean CAELEN**
*LCP Institut d'Electronique, USTHB, BP 32 El Alia-ALGER-ALGERIE
Tél: 2 75 94 57 - Fax:2 75 94 57
**CLIPS/IMAG, BP 53, 38041 Grenoble cedex 9 FRANCE
Tél: 76 51 46 27 - Fax: 76 44 66 75 - e-mail: Jean.Caelen@imag.fr

ABSTRACT

The original features of the arabic phonetic system are mainly based on the presence of emphatic and geminate consonants. Duration is a pertinent feature for vowels. In this paper we present experiments on arabic phone recognition using automatically derived indicative features. This process is conducted by SARPH, an ear model-based labelling system. The phone identification task is made by a finite state network (FSN) which follows the signal segmentation. Our interest goes to fricative consonants (geminate and emphatic) and vowels (long and brief).

1. INTRODUCTION

L'originalité de la phonétique arabe se fonde, pour une grande partie sur la pertinence de la durée dans le système vocalique et sur la présence de consonnes emphatiques. Une autre caractéristique déterminante est la gémination. Celle-ci joue un rôle fondamental dans le développement morphologique nominal et verbal. Dans cet article, il est question de la reconnaissance automatique par SARPH (Système Arabe de Reconnaissance par Phones) des voyelles longues et brèves, des fricatives emphatiques et non emphatiques ainsi que des fricatives géminées. La méthodologie suivie s'inspire de celle qu'utilise le décodeur acoustico-phonétique du système DIRA pour le français (Caelen, 1988). SARPH effectue tout d'abord une extraction d'un large éventail de paramètres acoustiques (F0, formants, indices acoustiques, énergie,...) pour réaliser ensuite une segmentation automatique en phones homogènes. L'étape d'identification est réalisée au moyen de réseaux d'états finis régis par des règles de production phonétiques.

2. OBJECTIF

Dans ce travail, nous tentons d'apporter notre contribution dans le débat qui partage les phonéticiens, linguistes et autres, sur l'emphase, la gémination et la pertinence de la durée pour les voyelles arabes. Nous proposons de déceler ces traits automatiquement au moyen d'indices acoustiques adéquats. Nous validerons cette démarche par SARPH en montrant son aptitude à reconnaitre les voyelles longues, les fricatives géminées et emphatiques.

3. DESCRIPTION PHONETIQUE DE LA LANGUE ARABE

3.1. La quantité vocalique

Le système vocalique comprend deux quantités phonologiques pour chaque timbre. A chaque voyelle brève /a/, /i/, /u/ s'oppose respectivement une voyelle longue /a:/, /i:/, /u:/. Pour l'arabe cette opposition temporelle brève/longue est fondamentale aux niveaux grammatical et sémantique. Par exemple, les deux mots /3amal/ "chameau" et 3ama:l "beauté" ne diffèrent que par la longueur de la voyelle finale. Dans la tradition des grammairiens arabes, une voyelle longue est ressentie comme deux voyelles brèves. Les travaux de Cantineau (Cantineau, 1960) qui constituent pour le vocalisme de l'arabe une référence indéniable ne vont pas plus loin. El Ghazali (Ghazeli, 1979) remet en question cette analyse traditionnelle par l'introduction de la notion de tension et de laxité pour les voyelles. Jakobson dans ses *preliminaries* (Jakobson, 1963) conclut qu'une voyelle tendue est plus longue que son homologue brève. Dans notre cas, la nécessité d'élaborer une taxinomie qui faciliterait la formalisation de règles d'identification utilisé par les systèmes de reconnaissance automatique de la

parole (les systèmes basés connaissances), nous a amené à tenir compte du trait tendu/lâche afin d'établir une opposition des voyelles brèves et de leurs correspondantes longues (Boudraa & Selouani, 1994). Le modèle calculatoire des traits acoustiques basé sur le modèle d'oreille (Caelen, 1979) est utilisé pour quantifier puis coder ce trait, ce qui permet une automatisation de la distinction voyelle longue / voyelle brève par SARPH.

3.2. L'emphase

Les consonnes emphatiques sont au nombre de quatre /t/, /s/, /d/, /ð/. Elles sont articulées dans la partie antérieure de la cavité buccale, la racine de la langue est reportée en arrière contre la paroi pharyngale postérieure et un creusement du dos de la langue est observé. Acoustiquement, elles se caractérisent par l'élévation de la transition de F1 et la baisse de la transition de F2 de la voyelle précédente et suivante. Dans notre système c'est au moyen de l'indice bémolisé / diésé que nous arrivons à une opposition emphatique / non emphatique (Boudraa & Selouani, 1994).

3.3. La gémination

Elle se manifeste par le renforcement de l'articulation et une prolongation de la fermeture de la plosive ou du continuant des autres consonnes. Là aussi, l'école traditionnaliste s'impose par le fait qu'elle considère la gémination comme un simple dédoublement de la consonne. S'il est évident qu'il existe une différence de durée notable entre la consonne géminée et son homologue simple (un rapport de 2), Bonnot (Bonnot, 1979) se garde de considérer, la durée comme seul principe de discrimination. Il est indispensable suggère-t-il, "d'accorder une très grande attention aux autres indices". Dans notre cas, nous préconisons l'introduction de l'indice tendu/lâche en plus du paramètre durée pour une meilleure distinction de la consonne géminée et de sa correspondante simple.

4. LE SYSTEME SARPH

Les différents traitements sont effectués dans des trames de 8 ms avec une cadence d'échantillonnage de 16 KHz. L'énergie ainsi que le taux de passage par zéro sont calculés. La fréquence fondamentale est estimée et corrigée par la technique de l'ambiguité modifiée. L'extraction des formants est effectuée sur le spectre LPC. A partir du modèle d'oreille de J.Caelen, les énergies de sortie de 24 filtres couplés correspondant à une portion de la membrane basilaire sont calculées. Celles-ci permettent de déterminer les indices acoustiques statiques que nous jugeons pertinents. Ils sont au nombre de 7: aigu/grave, fermé/ouvert, bémolisé/diésé, écarté/compact, doux/strident, continu/ discontinu, tendu / lâche.

4.1. La segmentation

Un codage delta des indices acoustiques est effectué afin de déceler leurs variations. Une fonction qui somme les sorties absolues des différents codeurs est estimée, quantifiant ainsi, la discontinuité entre deux trames successives. Si cette somme est supérieure à un seuil variable dans le temps, une marque est attachée à la trame courante. Les trames comprises entre deux marques successives constituent un phone homogène. Pour chaque phone, les valeurs moyenne, maximale et minimale de chaque paramètre sont calculées. Les indices acoustiques statiques et dynamiques sont codés (d'une manière non linéaire) en 5 niveaux: --,-,0,+,++. Une relation d'ordre existe entre ces 5 degrés de codage, ainsi, TN++ signifie très tendu et TN-- très lâche. D'autres paramètres acoustiques tels que l'indice de friction et l'indice vocalique, sont calculés. Ils constituent avec les autres paramètres, la base de connaissances de SARPH.

4.2. L'identification

A chaque macro-classe est associé un réseau phonétique représentant la connaissance sur la macro-classe. Celui-ci est appliqué indépendamment sur la suite de phones. Un réseau est constitué d'un ensemble d'états et d'un ensemble de transitions. Les états représentent toutes les réalisations possibles des différentes phases acoustiques des macro-classes phonétiques. A chaque transition (ou arc) on associe une liste de contraintes (règles) à vérifier, une liste d'actions à effectuer en cas

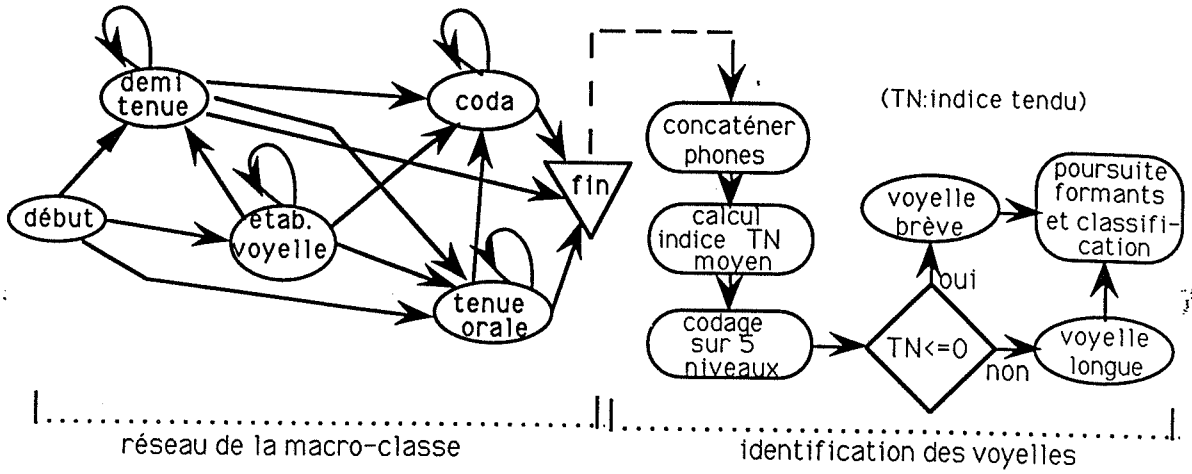


figure 1: Processus de reconnaissance des voyelles par SARPH.

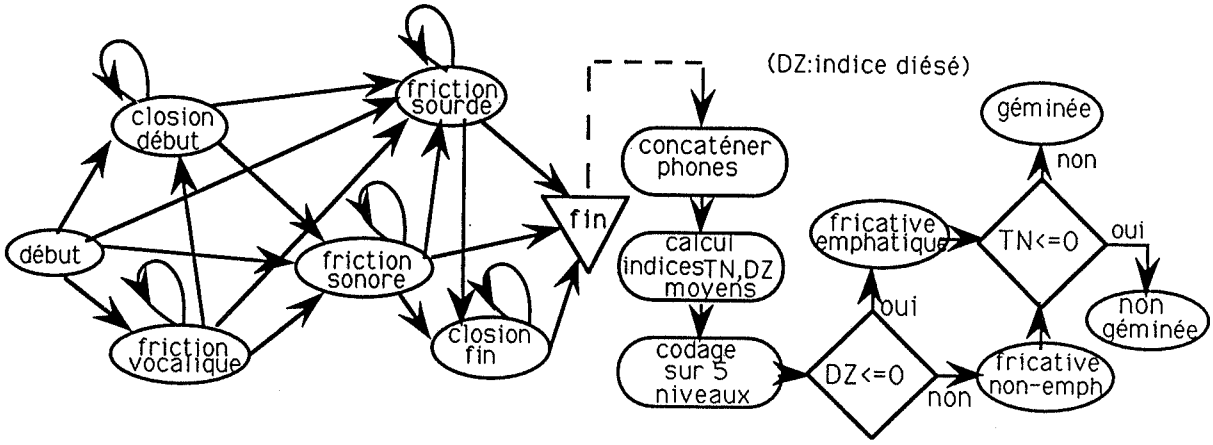


figure 2: Reconnaissance de la macro-classe fricative, du trait d'emphase et de la gémination.

de succès et enfin un score à chaque passage par la transition. Un phone peut être étiqueté par un ou plusieurs réseaux comme il peut être rejeté par tous. Le nombre de phones n'est pas connu au préalable et l'étiquetage d'un phone de rang N ne se fait que si les N-1 phones le précédant ont été étiquetés.

L'accès se fait donc séquentiellement et justifie l'utilisation d'une liste chaînée pour la modélisation du réseau phonétique. Ainsi chaque maillon de la liste représente un phone et contient les informations relatives à celui-ci. La liste linéaire chaînée est bidirectionnelle afin de permettre le retour arrière lors de l'exploration du réseau en profondeur. Les figures 1 et 2 shématisent le cheminement dans le réseau des voyelles et des fricatives de SARPH. Dans le cas des voyelles, si pour le phone courant , un 'établissement' est observé, on tentera d'associer au phone suivant la même phase ou bien les phases 'demie tenue' ou 'tenue orale' (notez l'absence de phase de nasalité car il n'existe pas de voyelles nasales en arabe). On opère de la même manière pour les phones suivants en tenant compte des passages permis dans le réseau et ce jusqu'à en sortir en cas de solution. Plusieurs étiquetages sont possibles grâce au processus de retour en arrière. Seule la solution présentant le score le plus élevé sera validée. Les règles du réseau sont de la forme:

SI (condition i) ALORS (action i) ET (incrémenter score_passage_réseau).

L'évaluation de la condition i renvoie à des calculs de prédicats et de fonctions.

5. RESULTATS ET DISCUSSION

Le corpus de test a été prononcé par 6 locuteurs (3 hommes et 3 femmes) algériens. Ces mêmes locuteurs ont participé à l'apprentissage. Les stimuli sont constitués de 40 occurences VCV et de 20 phrases (Arabe standard) où les fréquences d'apparition des phonèmes sont respectées (Mrayati, 1987). Le test concerne les 14 fricatives: /f/, /s/, /s̲/, /z/, /3/, /h/, /ħ/, /ʃ/, /ð/, /ð̲/, /χ/, /θ/, /γ/, /ɛ/.

Il faut cependant noter, que les autres consonnes (plosives, liquides, ...) figurent dans le corpus. Au total, le nombre de voyelles testées est de 852 et celui des fricatives est de 384.

Dans une première étape, SARPH a pour tâche de détecter les macro-classes. Les résultats obtenus sont présentés dans la table 1. La distinction voyelle longue-brève est réalisée avec succès dans 66 % des cas. Grâce à un algorithme de poursuite des formants (sur spectre LPC), SARPH opère une classification totale des voyelles avec un taux de 60 %. Les deux fricatives emphatiques sont décelées avec un taux de succès de 71% (résultats présentés dans la table 2). Une suite supplémentaire de 108 occurences VCV dont la consonne est une fricative géminée a été testée et le taux de détection correct de la gémination est de 64 % . Le choix de tenter la détection de la gémination indépendament des autres consonnes est délibéré, dans la mesure où phonétiquement il n'existe pas de consonnes géminées (les utiliser dans le corpus général déséquilibrera phonétiquement celui-ci).

6. CONCLUSION

A travers cette étude nous confirmons par l'expérience 3 faits concernant le système vocalique et consonantique de l'arabe, mis en relief avec plus ou moins d'insistance dans (Ghazeli, 1979), (Bonnot, 1979), (Jakobson, 1963) et (Boudraa & Selouani, 1994) à savoir :

-L'indice tendu/lâche permet d'arriver à une discrimination entre voyelle longue et brève.

-L'emphase est décelée efficacement au moyen de l'indice bémolisé/diésé.

-Une distinction entre la consonne géminée et son homologue simple par l'indice tendu/lâche, est possible.

Table 1: Taux de reconnaissance des macro-classes.

macro-classe	détectées	non détectées	taux
voyelles	817	35	96 %
fricatives	288	96	75 %

Table 2: Taux de reconnaissance des voyelles, de l'emphase et de la gémination.

phonème	reconnu	non reconnu	taux
voyelles a:, u:, i:	112	58	66%
voyelles a,u,i,a:,u:,i:	511	341	60 %
fricatives s̲ ð̲	28	11	71 %
géminées	69	39	64 %

7. BIBLIOGRAPHIE

Bonnot J.F. (1979) "Etude expérimentale de certains aspects de la gémination et de l'emphase en arabe", travaux de l'institut phonétique de Strasbourg N°11, pp 109-118.

Boudraa B., Selouani S.A (1994), "matrices phonétiques et matrices phonologiques arabes" XXèmes JEP, Tregastel .

Caelen J. (1979), "un modèle d'oreille, analyse de la parole continue, reconnaissance phonémique", thèse de doc. d'état. Sciences, Toulouse.

Caelen J., Tattegrain H. (1988), "Le décodeur acoustico-phonétique dans le projet DIRA", XIIèmes JEP, Nancy , pp 115-121.

Cantineau J. (1960) "Cours de phonétique arabe", Klincksieck, (1ere Ed.1941), Paris.

Ghazeli S. (1979) "Du statut des voyelles en arabe ", analyses-théories, études arabes, N°2-3, pp 199-219.

Jakobson R., Fant G.M., Halle M. (1963), "preliminaries to speech analysis: The distinctive features and their correlates", MIT press.

Mrayati M. (1987),"Statistical studies of arabic roots", Applied arabic linguistics and signal and information processing, Hamshire publishing.

JEP 96

RECONNAISSANCE AUDIOVISUELLE

AVIGNON 10-14 JUIN 1996

ASYNCHRONIE DANS LES SYSTÈMES DE RECONNAISSANCE DE LA PAROLE BASÉS SUR LES HMM

Pierre JOURLIN

Laboratoire d'Informatique - 339, Chemin des Meinajariès - BP 1228 - 84911 Avignon Cedex 9

Tél.: 90 84 35 35 - Fax: 90 84 35 01 - e-mail: jourlin@univ-avignon.fr

ABSTRACT

The asynchrony between the different phonatory organs is a phenomenon which is well-known in the speech production and perception domain. Few bimodal connected speech recognition systems have taken it into account. After having presented them, we define a new method, based on the product of two HMM. Finally, we test our method on a corpus of connected letters pronounced by two differents speakers.

1. INTRODUCTION

Durant cette dernière décennie, quelques systèmes de reconnaissance de la parole bimodale ont été réalisés [1], [2]. Bien que les phénomènes de désynchronisation entre les différents organes phonatoires aient été largement étudiés [3] dans les domaines de la perception et de la production de la parole, ils ont été plus rarement pris en considération dans le domaine de la reconnaissance automatique de la parole bimodale [4] [5] [6]. Nous allons tout d'abord décrire l'asynchronie pour les HMM et sa prise en compte dans quelques systèmes actuels. Nous montrerons ensuite comment ce phénomène peut être géré avec un produit de HMM. Enfin, cette approche est évaluée avec des systèmes de reconnaissance. Ces travaux s'inscrivent dans le cadre du projet AMIBE [7].

2. L'ASYNCHRONIE

Entre deux réalisations différentes d'une même unité phonétique, la mise en correspondance d'événements acoustiques et labiaux est temporellement différente [3]. C'est cette variabilité que nous appellons par la suite asynchronie. Celle-ci est facilement gérable avec des modèles de Markov cachés pour les mots isolés. Supposons que l'on se situe, dans un premier temps, dans le cadre restreint d'un traitement bimodal de la parole, tel que les suites d'observations A_1^T et V_1^T correspondent à un seul mot (et non pas une suite de mots). Ce mot fait partie d'un vocabulaire $\{W_1, \ldots, W_n\}$. À ces n mots sont associés les modèles acoustiques et labiaux $\{Ma_1, \ldots, Ma_n\}$ et $\{Ml_1, \ldots, Ml_n\}$. La solution recherchée est le mot $\mathcal{M}$ qui maximise la probabilité conjointe :

$$\mathcal{M} = \arg max_i P(Ma_i|A_1^T) \times P(Ml_i|V_1^T) \quad \text{avec} \quad i \in [1, n].$$

L'asynchronie n'apparaît pas ici comme un problème, puisque la probabilité associée à chacune des sources est calculée séparément. Si l'on utilise des HMM émettant des vecteurs composés en partie d'observations acoustiques et en partie d'observations labiales, il faudrait un corpus de taille plus importante pour assurer l'apprentissage correct de l'asynchronie. Or, à l'heure actuelle, personne ne dispose de tels corpus. De plus, dans le cas de la parole continue et en conservant cette méthode, il s'agirait de rechercher non plus le couple de modèles mais le couple de suites de modèles maximisant cette probabilité. La complexité résultant de ce problème étant ingérable, nous devons donc avoir recours à des solutions sous-optimales.

2.1. Le Décodage N-Best

Une première possibilité [5] consiste a réaliser un décodage acoustique proposant les N meilleures suites de mots possibles auxquelles sont associées leurs probabilités. Nous pouvons alors calculer pour chacune de ces suites de mots, celle qui a la plus forte probabilité conjointe. L'asynchronie est donc gérée au niveau de la phrase, mais de façon très incomplète. Considérons un système qui choisirait pour solution parmi l'ensemble E des N meilleures solutions acoustiques et l'ensemble E' des N' meilleures

solutions labiales, celle qui a la plus forte probabilité conjointe. Pour pouvoir affirmer que la solution proposée est la meilleure, il faut qu'elle appartienne à $E \cap E'$. En effet, si cette solution S_i de probabilité acoustique p_i et labiale p'_i, optimale sur $E \cup E'$ appartient à $E \cap E'$, alors toute solution S_j de probabilités p_j et p'_j extérieure à $E \cup E'$, vérifiera : $p_j \times p'_j < p_i \times p'_i$, la définition du système imposant : $p_j < p_i$ et $p'_j < p'i$. Dans le cas contraire, on ne peut affirmer que l'asynchronie a été gérée de façon optimale.

2.2. Les modèles maître-esclave

Une seconde possibilité est l'utilisation d'un modèle acoustique *piloté* par un modèle labial [4]. Pour passer d'un état du modèle acoustique à un autre, il existe autant de transitions qu'il y a d'états dans le modèle maître (labial). Le modèle ainsi créé doit être soumis à une phase d'apprentissage qui, étant donné le grand nombre de paramètres à estimer, demande un très grand nombre de données. En pratique, avec un corpus tel que celui d'AMIBE [7], il faudra simplifier énormément le modèle labial. De plus, l'asynchronie n'est gérée qu'à l'intérieur des unités de reconnaissance.

2.3. Le produit de modèles

Nous proposons un autre type de combinaison dont l'idée forte consiste à réaliser au décodage le simple produit d'un modèle acoustique et d'un modèle labial (voir figure 1).

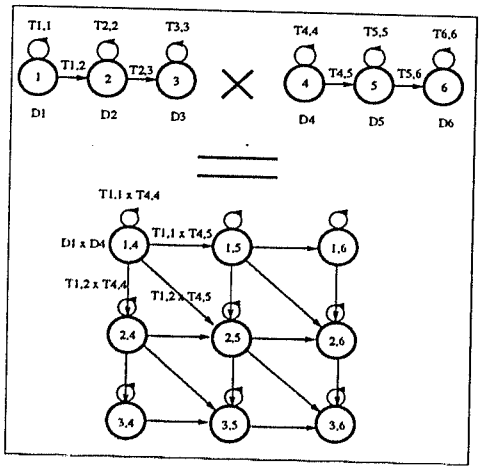


Figure 1: Exemple de produit

Le lecteur intéressé trouvera plus de détails en [8]. Pour une suite d'observations donnée, la probabilité calculée avec le modèle produit est égale au produit des probabilités données par le modèle acoustique et le modèle labial.

L'originalité de cette approche est donc qu'elle gère de façon optimale l'asynchronie à l'intérieur des unités de reconnaissance. Ceci nous permet d'évaluer les perturbations introduites par cette forme de variabilité dans des HMM de topologie classique. De plus, le produit est réalisé après un apprentissage séparé des deux modèles, ce qui nous permet d'éviter d'avoir recours à des modèles simples pour les données labiales. Les modèles ainsi obtenus sont utilisés dans l'algorithme classique du *token passing model* [9]. Dans son fonctionnement actuel, cet algorithme, comme dans le cas des modèles maître-esclave, ne permet de gérer l'asynchronie qu'à l'intérieur des unités de reconnaissance.

2.4. Le modèle synchrone

Nous décrivons ici les modèles synchrones classiques que nous utilisons pour nos différentes comparaisons [6]. Tous les modèles ont six états émetteurs, un état initial et un état final non-émetteurs. Les distributions associées sont des mixtures de deux gaussiennes.

Les modèles acoustiques émettent des vecteurs composés de douze coefficients MFCC et de l'énergie du signal auquels sont ajoutées leurs dérivées. Les modèles labiaux émettent des vecteurs composés de la hauteur, largeur, aire intero-labiale, de leurs vitesses et accélérations respectives. Les modèles acoustico-labiaux émettent la concaténation des vecteurs acoustiques et labiaux décrits ci-dessus. Il est à noter que les deux sources d'information sont pondérées, ceci afin de donner plus d'importance à la source acoustique. Les nouveaux modèles acoustico-labiaux seront construits par produit des précédents modèles acoustiques et labiaux après apprentissage séparé de ces derniers.

3. EXPÉRIMENTATIONS

Un corpus à été réalisé par l'Institut de la Communication Parlée de Grenoble [10] (locuteur : JLS), un autre par le Laboratoire d'Informatique d'Avignon (locuteur : PJ). Les deux corpus sont constitués de la prononciation des 26 lettres de l'alphabet français prononcées de façon continue. Pour chaque locuteur, les données sont composées de 200 enregistrements de 4 lettres, répartis en 70% pour l'apprentissage et 30% pour le test. Tous les systèmes précédemment décrits ont été réalisés et évalués à l'aide de la boîte à outils HTK du C.U.E.D.

3.1. Résultats

Table 1: Résultats pour le locuteur jls
It : Itération, A : Acoustique, L : Labial,
ALS : Bimodal synchrone, ALP : Bimodal produit

It	A	L	ALS	ALP
0	88.6	36.3	89.7	90.1
2	90.5	38.6	91.6	92.8
4	90.9	40.1	91.2	93.1
6	91.2	40.1	90.5	92.8
8	91.2	39.0	91.2	93.5
10	90.9	39.3	91.6	93.5
12	90.9	38.2	92.0	93.5
14	90.9	39.0	92.0	93.5
16	90.5	38.2	92.0	93.1
18	90.5	38.2	92.0	93.1

Table 2: Résultats pour le locuteur pj

It	A	L	ALS	ALP
0	73.7	21.9	75.8	76.7
2	73.2	24.5	76.7	78.0
4	74.5	24.1	77.1	77.1
6	75.0	24.5	77.5	76.2
8	74.1	24.5	76.7	76.7
10	73.7	25.8	76.2	76.2
12	73.2	25.8	77.1	76.2
14	70.6	25.4	77.1	76.7
16	70.6	27.5	77.5	75.8
18	70.6	28.0	77.5	76.7

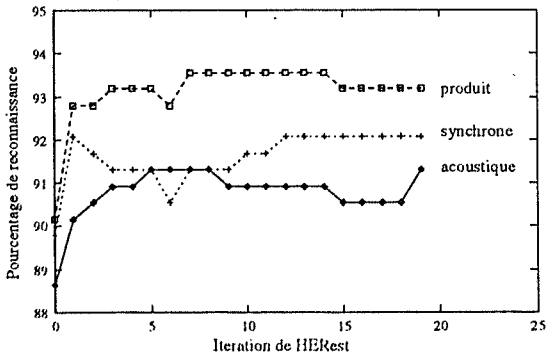


Figure 2: Résultats pour le locuteur jls

Pour chaque locuteur et pour chaque type de modèle, nous avons effectué une série de tests. Les modèles sont évalués pour chaque itération d'aprentissage, ceci afin de mettre en évidence

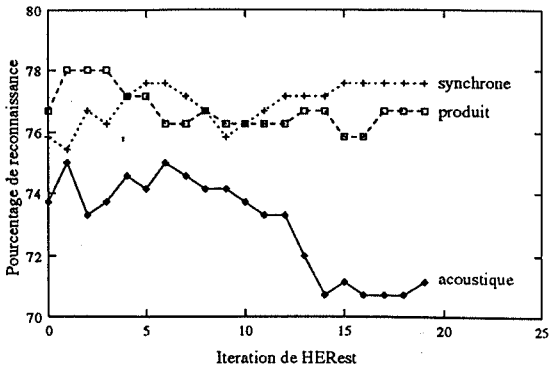


Figure 3: Résultats pour le locuteur pj

la variabilité des résultats, due au manque de données. Enfin, pour une itération donnée, le modèle acoustico-labial produit est construit à partir des modèles acoustiques et labiaux entrainés jusqu'à ce niveau. Les résultats (pourcentage de lettres reconnues, insertions déduites) sont récapitulés dans les tableaux 1 et 2 donnés ci-dessus. Ces résultats font apparaître une amélioration du modèle produit sur le modèle synchrone pour le locuteur jls, mais aucune amélioration significative pour le locuteur pj. Ceci semble indiquer une différence de quantité d'asynchronie présente pour chaque locuteur. Il faut néanmoins se garder, en l'absence de corpus suffisant, d'attacher trop d'importance aux résultats dans leur valeur absolue.

4. CONCLUSION

Nous avons montré que l'on peut tirer parti d'une gestion des phénomènes d'anticipation et de rétention dans un système fondé sur les modèles de Markov cachés. En outre, la méthode que nous avons proposée permet la fusion de deux modèles de topologies différentes. Nous pouvons donc choisir la topologie optimale pour chaque partie du vecteur d'information et pour chaque unité de reconnaissance avant de les fusionner. D'autre part, la phase d'apprentissage étant appliquée séparément aux modèles acoustiques et visuels, cette méthode ne nécessite pas un grand nombre de données et ne pose pas de problème de complexité pour cette phase, contrairement au modèle maître-esclave [11]. Enfin, elle ne requiert aucune modification des algorithmes classiques utilisés et peut donc être appliquée à n'importe quel système probabiliste.

5. PERSPECTIVES

Une première voie d'amélioration consisterait à gérer ces phénomènes également à l'extérieur des frontières des unités de reconnaissance. Dans ce but, il sera nécessaire d'élaborer une stratégie de décodage spécifique. D'autre part, dans le contexte de la parole unimodale, il serait possible de tirer profit des modèles produits à la condition de procéder au préalable à une inversion de modèle articulatoire sur les paramètres acoustiques [12], ceci dans le but d'obtenir des paramètres articulatoires entre lesquels apparaît l' asynchronie. Cependant, bien que la recherche dans ce domaine soit en progression, des travaux supplémentaires doivent être réalisés en vue du traitement de la parole continue. La combinaison de modèles que nous avons définie nous permet d'avoir des modèles efficaces en termes de reconnaissance, mais qui ont un nombre d'états pouvant être très grand. Il serait donc également intéressant d'étudier des modèles non optimaux plus efficaces en terme de rapidité. Enfin, la fiabilité des résultats obtenus en reconnaissance de la parole bimodale est assez faible et il est difficile de montrer la supériorité d'une modélisation par rapport à une autre. En effet, les techniques d'acquisition et d'extraction de données visuelles en temps réel ne permettent pas encore d'obtenir des corpus de taille équivalente à celles des corpus unimodaux. Cependant, on peut espérer que l'engouement des chercheurs en parole pour ce domaine accélère la mise au point de ces nouveaux corpus.

6. BIBLIOGRAPHIE

[1] E.D. Petajan (1984).
Automatic lipreading to enhance speech recognition
Proceedings of the Global Communications Conference, IEEE Communication Society, Atlanta, Georgia, 265-272.

[2] J. Robert-Ribes (1995)
Modèles d'intégration audio-visuelle de signaux linguistiques : de la perception humaine à la reconnaissance automatique de voyelles
Thèse de doctorat - ICP Grenoble

[3] C. Abry, M.T. Lallouache (1994)
Pour un modèle d'anticipation dépendant du locuteur - Données sur l'arrondissement en français
Bulletin de la communication parlée - ICP Grenoble

[4] B. Jacob, C. Sénac, R. André-Obrecht, F. Pellegrino (1995)
Improving Speech Recognition With Multimodal Articulatory Acoustic HMMs
Proceedings of International Congress of Phonetic Sciences, Stockholm, vol. 4, 284-287

[5] A. Foucault (1995)
Système acoustico-labial de reconnaissance de la parole Rencontres Jeunes Chercheurs en Parole, Paris

[6] P. Jourlin, M. El-Bèze, H. Méloni (1995)
Integrating visual and acoustic informations in a speech recognition system based on HMM
International Congress of Phonetic Sciences, Stockholm, vol. 4, 288-291

[7] C. Montacié, M.J. Caraty, R. André Obrecht, L.J Boë, P.Deléglise, M. El-Bèze, I. Herlin, P. Jourlin, T. Lallouache, B. Leroy, H. Méloni (1995)
Applications Multimodales pour Bornes et Interfaces Multivaluées Ecole Thématique : Traitement automatique de la parole : Fondements et Perspectives, Marseille, 155-164

[8] P. Jourlin (1996)
Handling Disynchronization Phenomena with HMM in Connected Speech
European Signal Processing Conference, Trieste (accepté)

[9] S.J. Young, N.H. Russel, J.H.S. Thornton (1989)
Token Passing : a Simple Conceptual Model for Connected Speech Recognition Systems
Technical Report - CUED/F-INFENG/TR.38

[10] M.T. Lallouache (1991)
Un poste "visage-parole" couleur
Thèse de doctorat - INPG Grenoble

[11] F. Brugnara, R. De Mori, D. Giuliani, M. Omologo (1991)
A parallel HMM approach to speech recognition
Proceedings of Eurospeech'91, Gènes, 1103-1106

[12] L. Candille, H. Méloni (1995)
Automatic speech recognition using production models
International Congress of Phonetic Sciences, Stockholm, vol. 4, 256-259

DETECTION ET LOCALISATION AUDITIVE ET VISUELLE D'EXPLOSIONS CONSONANTIQUES DANS DES SEQUENCES VCV BRUITEES

M. PIQUEMAL, J.L. SCHWARTZ, F. BERTHOMMIER, T. LALLOUACHE, P. ESCUDIER

ICP, CNRS URA 368, INPG – Université Stendhal 46 av Félix Vialet, 38031 Grenoble cedex 1, France

Tél.: 76 57 47 15 - Fax: 76 57 47 10 - e-mail: piquemal@icp.grenet.fr

ABSTRACT

This work presents an original attempt to define a suitable architecture for the temporal processing of speech by ear and eye, based on both physiological and functional constraints. First, an "auditory" system detects the location of a plosive release (burst) in VPV sequences (with V a vowel and P a plosive) from the acoustic signal: we show that its design provides a strong robustness even with white noise perturbations. Good correlates were also found with the second "visual" system charged with detecting the same events from lips geometry parameters. These two detectors provide a first representation of timing in VPV sequences, and seem to be a crucial step in order to better understand and exploit the auditory-visual sensor fusion in speech perception and recognition.

1. INTRODUCTION: LE 'MODELE TIMING-CIBLES'

Afin de définir et d'implanter une architecture plausible d'un point de vue biologique et fonctionnel, et efficace pour le traitement de la parole, il faut prendre en compte trois de ses principales propriétés: (1) le message est transmis par des stimuli intrinsèquement *dynamiques,* (2) ces stimuli sont *à la fois auditifs et visuels* (les malentendants peuvent lire sur les lèvres, et la lecture labiale peut être cruciale en environnement bruité), et enfin, (3) les stimuli de parole audiovisuelle sont produits par le *système moteur humain*, avec ses propres contraintes. En résumé, la parole est dynamique, audiovisuelle et motrice.

Les caractéristiques dynamiques ont conduit Chistovich (Chistovich, 1980) à proposer à la fin des années 70 un modèle fondé sur deux systèmes parallèles, mais non indépendants, pour l'analyse auditive: un premier chargé de la détection des événements acoustiques dans le domaine temporel, et un second, contrôlé par le premier, donnant continûment un certain nombre de représentations spectrales du signal d'entrée. L'intérêt de ce modèle est qu'il sépare le traitement auditif de l'information temporelle et spectrale dans deux sous-systèmes différents, chacun ayant des caractéristiques temps-fréquence spécifiques optimisées pour sa tâche. Nous avons généralisé cette architecture en proposant le 'Modèle Timing-Cible' pour la perception audiovisuelle de la parole. Dans ce modèle, on considère deux sous-systèmes séparés pour le traitement audiovisuel (AV) de la parole; l'un chargé de détecter les événements temporels caractérisant le timing moteur, et l'autre suivant les mouvements articulatoires pour en déterminer les cibles spatiales (Schwartz et al., 1992). Ces sous-systèmes s'appellent respectivement système 'Phasique AV' (ou de 'détection d'événements') et système 'Tonique AV' (ou 'd'analyse des trajectoires').

Nous avons déjà effectué de nombreux travaux sur le système tonique AV (Robert-Ribes et al., in press). Dans cet article, nous nous focaliserons sur le système phasique AV. Cependant, avant de développer un système phasique audiovisuel complet, il nous semble essentiel de mieux comprendre la compatibilité et la complémentarité de chaque sous-système: auditif et visuel. Tel est le propos de cette étude dans laquelle nous tentons de résoudre le problème classique et fondamental de la détection de l'explosion consonantique (burst) dans des séquences Voyelle-Plosive-Voyelle.

2. CORPUS

Il est constitué de séquences $V_1PV_2PV_1$, où P est une plosive parmi /b, d, g/ et V_i une des quatre voyelles canoniques /i, y, u, a/. Ces séquences sont prononcées trois fois chacune par un seul locuteur, avec donc 96 occurences de chaque plosive ($4xV_1$, $4xV_2$, 3 répétitions, 2 occurences par séquence $V_1PV_2PV_1$). Nous avons également enregistré, en guise de contrôle, un corpus de séquences $V_1V_2V_1$

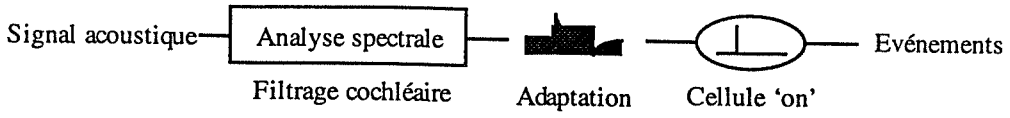


Figure 1: Architecture du détecteur d'événements auditifs.

(mêmes voyelles) et un corpus de doubles plosives dont nous ne traiterons pas ici.

Les données vidéo synchrones sont acquises et enregistrées avec le signal audio par un magnétoscope. Ensuite, un système d'analyse d'images (Lallouache et al.,1988) mesure sur chaque trame vidéo les paramètres géométriques des lèvres A (ouverture longitudinale), B (ouverture verticale) and S (aire intéro-labiale). On obtient, ainsi, pour chaque paramètre, son évolution temporelle.

3. LE SYSTEME AUDITIF PHASIQUE

L'architecture du système auditif phasique repose sur trois modules de base, respectivement (1) l'analyse spectrale dans la cochlée (2) le renforcement des variations d'énergie locales par l'adaptation nerveuse dans le nerf auditif, et (3) la détection d'événements par des cellules 'on' du système nerveux central.

3.1 Bases physiologiques

3.1.1. Analyse cochléaire

Tous les modèles de traitements auditifs commencent par cette première étape de traitements, produisant la base de la géométrie des sons: la représentation 'tonotopique'.

3.1.2. Adaptation nerveuse dans les cellules primaires

L'adaptation est une propriété caractéristique de la réponse de chaque fibre du nerf auditif qui accentue les modulations rapides d'amplitude que l'on trouve dans toutes les langues, et ainsi joue un rôle important dans le traitement de la parole dans le nerf auditif (Delgutte, 1986).

3.1.3. Regroupement et détection dans les cellules secondaires 'on'

L'estimation du timing se produit grâce à un système capable de détecter des événements acoustiques. Le noyaux cochléaire est pourvu d'unités biologiques qui semblent pouvoir effectuer de telles tâches: les neurones 'on' (Young, 1984). Il est cependant probable que ces cellules signalent les modulations rapides associées au pitch, et qu'il faille pénétrer plus avant dans le système nerveux central pour trouver, au niveau du colliculus inférieur ou du cortex auditif primaire par exemple, des unités 'on' qui détectent des événements basse fréquence comme les explosions consonantiques (Delgutte, 1996). L' ICP a déjà proposé un premier modèle fonctionnel de ces cellules 'on' (voir aussi Delgutte, 1986). Le fonctionnement du modèle repose sur (i) une forte réponse à un événement acoustique dans une bande de fréquence (canal), et (ii) une sommation des réponses selon l'état de synchronie de différents canaux (Wu et al., to appear).

3.2 Implémentation des modèles

Le logiciel développé pour la détection d'événements acoustiques, reprend de très près les 3 points de l'architecture décrite ci-dessus: on y retrouve les principales composantes physiologiques (figure 1). Cependant, tout en essayant de garder des principes algorithmiques les plus proches possibles des données biologiques, nous avons voulu utiliser des méthodes standard de traitement du signal afin de pouvoir intégrer nos systèmes à n'importe quelle plate-forme.

3.2.1. Analyse spectrale

Le signal acoustique est échantillonné à 16 kHz sur une dynamique de 16 bits. Après une pré-emphase classique par un filtre de 6 dB/oct, l'analyse spectrale est calculée sur des fenêtres de 5 ms décalées de 0.5 ms entre elles: à ce niveau nous avons donc un échantillonnage de 0.5 ms.

Nous avons simulé le concept classique de bandes critiques en utilisant la méthode du 'warping' non linéaire en fréquence proposé par (Oppenheim et al., 1972). La sortie de cette analyse spectrale nous fournit donc un spectre sur 64 canaux.

3.2.2. Adaptation

Le modèle que nous utilisons a été développé par Wu (1990). Les phénomènes d'adaptation rapide et à court terme sont considérés comme deux processus indépendants, modélisés par des filtres du premier ordre. Chacun des filtres comporte une non-linéarité saturante et dispose de constantes de temps et de dynamiques différentes. Ce modèle a montré ses capacités à bien rendre compte de la plupart des propriétés des réponses des fibres du nerf auditif.

3.2.3. Cellules 'on''

Le modèle de cellules 'on' consiste en deux parties: d'abord une intégration spectrale large bande suivie d'un traitement temporel renforçant des maxima locaux d'une durée de 10 à 20 ms, durée typique du burst de plosives.

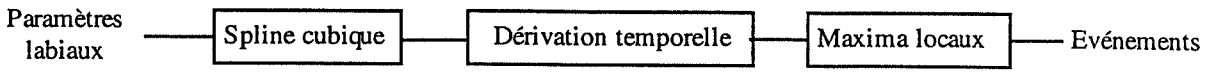


Figure 2: Architecture du détecteur d'événements visuels.

- Intégration spectrale

Les sorties des 'cellules primaires' correspondant aux bandes de 0.1 kHz à 2.5 kHz sont sommées. Ceci augmente l'immunité au bruit dans le cas d'une explosion large bande superposée à du bruit incohérent.

- Traitement temporel

Son rôle est de rendre compte de la caractéristique de base des cellules 'on', à savoir un très important renforcement d'augmentations locales d'énergie. Le modèle choisi, est un filtre passe bande (Berthommier, 1992). Remarquons que sa réponse temporelle est bien adaptée aux caractéristiques temporelles des bursts. Ainsi, la réponse impulsionnelle est positive pendant les 12 premières ms, puis la période d'inhibition (activité en dessous de 0) s'étend pendant une durée typique de syllabe: l'activité retrouve une valeur proche de 0 (10 % de la réponse négative maximum) 70 ms après le début de l'activation. Ceci filtre en quelque sorte la voyelle suivant l'explosion, et prépare le système pour le prochain burst.

3.2.4. Post-traitement

A la sortie de notre unique cellule 'on', nous avons un signal continu duquel nous devons extraire les événements acoustiques, et plus particulièrement dans notre cas présent, les explosions de plosives en contexte VPV. Ceci est fait par un traitement spécifique comportant deux étapes:

- Compression sygmoïdale

Cette compression de dynamique entre 0 et 1 est obtenue par seuillage de la sortie de la cellule 'on'.

-Détection et filtrage des événements supplémentaires

La détection proprement dite est effectuée par simple seuillage. Cependant les événements parasites sont éliminés par des règles ad-hoc (Piquemal et al., 1994). L'idée est d'éliminer les cas où, localement, le seuil est franchi plusieurs fois, comme par exemple lors de fluctuations d'amplitude dans une voyelle. L'événement est défini comme étant le premier maximum local suivant, dans les 5 ms, le passage d'une valeur seuil. Ensuite, si deux événements sont détectés dans un intervalle de 20 ms, seul le premier est retenu. Enfin afin de filtrer les fausses alarmes dans le bruit (détections au sein d'une voyelle), un critère sur le rapport entre les niveaux moyens d'énergie avant et après l'événement candidat a été ajouté.

4. SYSTEME PHASIQUE VISUEL

Des résultats préliminaires concernent les plosives bilabiales pour lesquelles nous n'avons pas besoin de traitements trop sophistiqués. La détection est basée sur les variations de l'ouverture longitudinale des lèvres. La détection se fait en trois temps (figure 2): lissage des trajectoire par 'spline' cubique, dérivation temporelle et détection du premier maximum local suivant une fermeture complète.

5. COMPATIBILITE ET COMPLEMENTARITE DES SYSTEMES PHASIQUES AUDITIFS ET VISUELS.

5.1 Comparaison des résultats de détections acoustiques sans bruit et visuelles

Nous avons analysé l'ensemble du corpus de plosives simples (3x96=288 explosions à détecter) par le module auditif phasique, et systématiquement vérifié (sur le signal et le spectrogramme associé) la validité de l'estimation. Nous avons également soumis le corpus de séquences $V_1V_2V_1$ au même module, afin de contrôler l'existence de fausses alarmes. Les résultats sont fournis dans le tableau I. Ils montrent une bonne performance d'ensemble du module (aucune fausse alarme dans les séquences vocaliques, 2 dans les séquences consonantiques; 96% de détections correctes des explosions consonantiques).

Nous avons également analysé le corpus de bilabiales par le module visuel phasique, avec 98% de détections correctes et aucune fausse alarme (table 1). Enfin, les détections fournies par les 2 modules pour les bilabiales sont temporellement très proches (moyenne -0.6 ms, écart type de 8 ms; voir table 2).

5.2 Dégradation, avec le bruit, du système acoustique.

Si l'on considère le nombre de détections manquées, et de fausses alarmes lors de la détection de bursts dans des stimuli bruités, on remarque que les résultats restent assez satisfaisants jusqu'à un niveau de bruit élevé (S/N = 13 dB), puis se dégradent nettement (S/N = 6 dB, niveau qui rend la plosive difficilement audible). De plus, lorsque la détection a lieu, celle-ci se produit de plus en plus tard au fur et à mesure que le niveau de bruit augmente (table 2).

Corpus	Détecteur	Détections à réaliser (Nbre)	Détections correctes (%)	Fausses alarmes (Nbre)
Voyelles sans bruit	Audio	0		0
Plosives sans bruit	Audio	288	96	2
Plosives 13dB	Audio	288	90	17
Plosives 6dB	Audio	288	79	52
Bilabiales	Vidéo	96	98	0

Table 1 : Résultats d'ensemble des modules de détections audio et visuel.

Détections	Décalage moyen
Bilabiales (Vidéo)	-0,6 ms
Plosives 13dB	2,5 ms
Plosives 6dB	9,4 ms

Table 2 : Positions respectives des localisations temporelles (référence: condition audio non bruitée)

6. CONCLUSION

Nous disposons maintenant de deux systèmes (acoustique et optique) capables de détecter de manière assez fiable des événements dans des séquences VPV, même lorsque celles-ci sont bruitées. La prochaine étape sera de fusionner ces donner afin de tirer le meilleur parti de ces deux modalités.

Nous devons garder en tête que la fusion audiovisuelle peut non seulement converger vers un même percept mais aussi en créer de nouveaux (McGurk, 1976). La stratégie que nous proposons pour la fusion des événements est la suivante:

- Premièrement attacher une probabilité à chaque position proposée par chaque détecteur (le ratio énérgétique évoqué lors de la sélection des candidats par le processeur audio semble être un bon indicateur).

- Ensuite, proposer des règles pour la fusion proprement dite: deux événements séparés par un certain délai (à préciser grâce à nos données enregistrées sur des plosives doubles) doivent être acceptés et considérés comme distincts; deux événements proches fusionnent en un seul.

Enfin, l'étape suivante sera de proposer deux systèmes toniques et leur mode de fusion: de précédents travaux (Robert-Ribes et all, in press) ont montré que la représentation motrice fournissait un cadre de travail intéressant pour la fusion. La fusion des stimuli auditifs et visuels dans leurs composantes tant phasiques que toniques devrait permettre l'élaboration d'un système de reconnaissance de plosives efficace, robuste dans le bruit, et compatible avec les performances humaines.

7. REMERCIEMENTS:

Nous tenons à remercier les projets Speech-MAPS (CEE ESPRIT-BR 6975) et Sphere (CEE CHCM-CT 93.0098) ainsi que la Région Rhône-Alpes (bourse de recherche pour M. Piquemal en relation avec la société ASTER) pour leurs soutiens.

8. REFERENCES:

Berthommier F. (1992): Intégration neuronale dans le système auditif, *Doctorat de l'Université Joseph Fourier.*

Chistovich L. (1980): Auditory processing of speech, *Langage and Speech*, Vol. 23, 67-73.

Delgutte B. (1986): Analysis of French stop consonants with a model of the peripheral auditory system, *Invariance and Variability of Speech Processes*, 163-177.

Delgutte B. (1996): Auditory Neural Processing of Speech. *The Handbook of Phonetic Science.*

Lallouache M.T.,Worley C. (1988): Saisie, édition et traitement d'images et signaux articulatoires: lèvres et mâchoire. *Journal d'Acoustique*, Vol. 1, 215-220.

McGurk H., MacDonald J. (1976): Hearing lips and seeing voices. *Nature*, 264, 746-748.

Oppenheim A.V., Johnson D.H. (1972): Discrete representation of signals. *Proc. of the IEEE*, 60, 681-691.

Piquemal M., Schwartz J.L. (1994): Towards an audiovisual "phasic" system: auditory and visual event detection in VCV sequences., *Speech Maps report 2* , WP4, 16-25.

Robert-Ribes J., Schwartz J.L., Escudier P. (in press): Fusion of the auditory and visual sensors in speech perception, *AI Review – Special Vol: Integration of Natural Language and Vision..*

Schwartz J.L., Arrouas Y., Beautemps D., Escudier P. (1992): Auditory analysis of speech gestures. *The Auditory Processing of Speech - From Sounds to Words* , 239-252.

Wu Z.L. (1990): Peut-on 'entendre' des événements articulatoires?, *Doctorat de l'INPG*

Wu Z.L., Schwartz J.L., Escudier P. (to appear): Physiologically-plausible modules and detection of articulatory-based acoustic events, *Advances in Speech, Hearing and Language Processing* ,Vol. 3.

Young E.D. (1984): Response characteristics of neurons of the cochlear nuclei, *Hearing Science, Recent Advances*, 423-460.

INTÉGRATION ASYNCHRONE DES INFORMATIONS AUDITIVES ET VISUELLES DANS UN SYSTÈME DE RECONNAISSANCE DE LA PAROLE

Alexandrina ROGOZAN, Paul DELÉGLISE, Mamoun ALISSALI

Laboratoire d'Informatique de l'Université du Maine - Avenue Olivier Messiaen - BP 535 - 72017 Le Mans Cedex
Tél. : 43 83 37 70 - Fax : 43 83 35 65 - e-mail : afoucaul@lium.univ-lemans.fr

ABSTRACT

This paper presents our work on the integration of visual data in an automatic continuous speech recognition system, based on Continous Hidden Markov Models.

We particulary aim at solving two problems:

- classification differences for the modeling of acoustic information (phonemes) and visual information (visemes) ;

- the phenomenon of anticipation and retention of visemes on the corresponding phonemes.

For this purpose we developed and tested three different systems. In order to evaluate the importance of each problem and to validate the integration model, we compare system performances. The comparisons show that some of the solutions we propose give satisfactory results, and suggest that further work on some others would lead to more performance improvement.

1. INTRODUCTION

Cet article présente les travaux que nous avons effectués sur l'intégration des informations visuelles dans un système de reconnaissance acoustique basé sur les Modèles de Markov Cachés Continus.

De nombreuses études (Summerfield, 1987) ont montré que la communication parlée est bimodale : les interlocuteurs utilisant non seulement des informations auditives mais aussi des informations visuelles. L'importance de ces dernières pour l'identification d'un message oral dépend de la dégradation du signal acoustique et de la complexité linguistique du message. Pour expliquer le comportement humain, différents modèles d'intégration des informations auditives et visuelles ont été proposés (Robert-Ribes, 1995).

De même, l'intégration des informations visuelles dans un système de reconnaissance automatique devrait permettre d'augmenter ses performances, plus particulièrement en milieu bruité. Mais pour cette intégration les modèles perceptifs ne sont pas, à l'heure actuelle, directement transposables. De nombreux problèmes sont à prendre en compte : en particulier les études articulatoires prouvent l'existence de sosies labiaux (Benoît *et al.*, 1994) et de phénomènes de rétention et d'anticipation labiale (Abry *et al.*, 1991). Nous qualifions ces phénomènes d'asynchrones en référence aux différences temporelles des mouvements des articulateurs. Cet asynchronisme nécessite un traitement différent du flux acoustique et du flux labial (problème noté P1). Les sosies labiaux imposent la distinction entre deux classes d'unités de reconnaissance : phonèmes pour les paramètres acoustiques, visèmes pour les paramètres labiaux (problème noté P2).

Pour des raisons techniques il existe une différence entre les périodes d'acquisition des vecteurs acoustiques et labiaux. Il est difficile d'atteindre la fréquence d'acquisition des paramètres acoustiques pour l'extraction des paramètres visuels (problème noté P3).

Certains systèmes de reconnaissance acoustico-labiaux existants proposent un modèle d'identification directe par concaténation des informations acoustiques et visuelles (Jourlin *et al.*, 1995). D'autres utilisent une identification séparée des entrées auditives et visuelles, suivie d'une fusion des résultats (Bregler *et al.*, 1993), (Adjouani *et al.*, 1995). Un dernier type de systèmes est basé sur deux Modèles de Markov Cachés (acoustique et

visuel) parallèles liés par une relation maître-esclave (Jacob *et al.*, 1995).

Nous avons développé trois systèmes pour tenter de mesurer l'apport des informations visuelles et l'importance relative de ces trois problèmes.

Ces systèmes sont construits à partir d'un système acoustique de référence (noté *S0*) :

- le premier système (*S1*) fusionne les informations acoustiques et visuelles au niveau des paramètres ;

- le deuxième (*S2*) est composé de deux sous-systèmes (acoustique et visuel), qui sont liés en contraignant le décodage au niveau visuel par les résultats du décodage au niveau acoustique ;

- le troisième (*S3*) est identique au précédent, mais il utilise une classification par visèmes pour le sous-système labial.

2. ARCHITECTURES DES SYSTÈMES DE RECONNAISSANCE

Le système de référence *S0* utilise des MMCC (Modèles de Markov Cachés Continus) acoustiques au niveau du phonème. L'analyse du signal est centiseconde, un vecteur d'observation étant fourni toutes les 10 ms. Le vecteur d'observation est constitué de 12 coefficients cepstraux à l'échelle Mel et de l'énergie totale de la fenêtre, auxquels sont ajoutées leur vitesse et leur accélération. Ce système, ainsi que les suivants, utilise un modèle de durée de la réalisation de l'unité de reconnaissance (phonème ou visème) (Suaudeau *et al.*, 1993).

L'information visuelle intégrée dans les systèmes acoustico-labiaux concerne le contour intérieur des lèvres du locuteur, représenté par trois paramètres : l'étirement labial, la séparation labiale et l'aire intéro-labiale (Lallouache *et al.*, 1990). Cette dernière est fortement corrélée aux produit des deux premiers paramètres, mais apporte néanmoins un supplément d'information pour les MMCC utilisés.

Dans le système *S1*, nous avons fusionné les informations acoustiques et les informations labiales. La période d'acquisition des paramètres labiaux (20 ms) est supérieure à celle nécessaire pour extraire les paramètres acoustiques (10 ms). Pour traiter ce problème sans perdre de précision dans le domaine acoustique, nous avons interpolé les paramètres labiaux à l'aide de fonctions spline sous tension (Cline, 1974). Le vecteur d'observation est donc constitué des paramètres fusionnés, ainsi que de leur vitesse et leur accélération. Ce vecteur est divisé en un flux acoustique et un flux labial pour permettre la pondération des deux modalités selon le niveau de bruit du signal de parole.

Cependant ce système ne permet pas le traitement des phénomènes de rétention et d'anticipation labiale. Pour tenir compte des ces phénomènes, il faut non seulement enlever la simultanéité des deux flux au niveau intra-modèle (comme le fait (Jourlin, 1995)), mais aussi au niveau inter-modèle.

Pour réaliser ces deux objectifs, nous proposons un deuxième système *S2*, constitué de deux sous-systèmes : acoustique et visuel. Le premier sous-système utilise des MMCC acoustiques, pendant que le deuxième sous-système utilise des MMCC labiaux (figure 1).

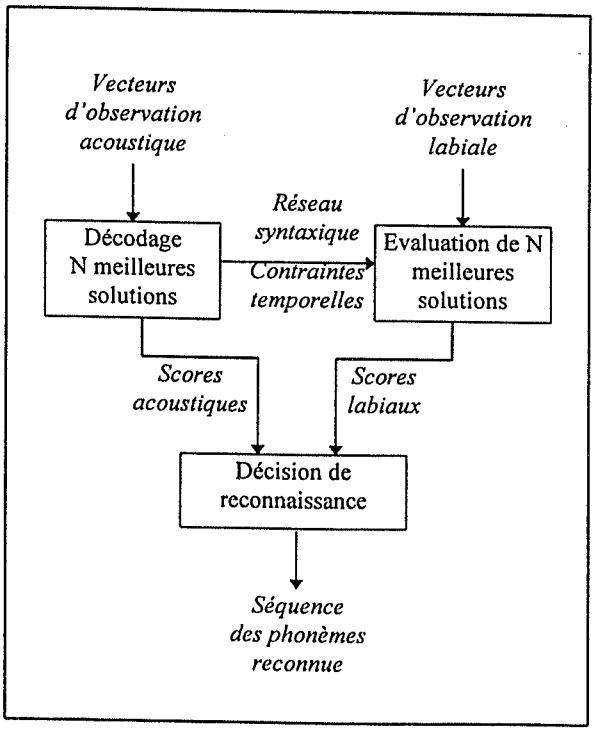


Figure 1 : Architecture du système à composantes acoustique et labiale séparées (S2)

L'apprentissage est réalisé séparément pour chacun des sous-systèmes à partir d'un

étiquetage phonétique du corpus, réalisé au niveau acoustique. La première étape consiste en l'obtention des N meilleures solutions de reconnaissance phonétiquement différentes à l'aide du premier sous-système.

Ces solutions sont ensuite utilisées pour construire un réseau syntaxique augmenté de contraintes temporelles, déterminées à partir des informations sur les frontières de phonèmes. Ces contraintes fixent les limites dans lesquelles le sous-système labial peut poser les frontières phonémiques. Un score labial est ainsi calculé pour chacune des solutions précédentes.

Le score final est calculé par combinaison linéaire du score acoustique et du score labial. Le coefficient de pondération est déterminé (pour les expériences présentées) de façon empirique, sa valeur étant en rapport avec le niveau de bruit. À titre d'exemple, la valeur du coefficient de pondération du flux labial est de 0,33 à -10 dB et de 0,22 sans bruit. La décision de reconnaissance consiste à choisir la solution ayant le meilleur score final.

L'utilisation de la même unité (le phonème) pour caractériser les deux composantes (acoustique et labiale) du système *S2* ne permet pas de tenir compte des sosies labiaux.

Par contre le système *S3* (qui est une variante du précédent) utilise le visème, au lieu du phonème, comme unité de décision pour la composante labiale. Les visèmes représentent l'ensemble des unités distinctives et indissociables qui décrivent le signal de parole sur le plan visuel. L'apprentissage de la composante labiale est ainsi rendu plus cohérent par le regroupement des phonèmes en classes en fonction de leur similarité visuelle (Ficher, 1968).

Après le décodage au niveau acoustique, la séquence de phonèmes propre à chacune des N meilleures solutions est transcodée en une séquence de visèmes, avant d'être prise en compte dans le réseau syntaxique. Ce qui permet d'affecter un score labial à cette séquence et de déterminer, de manière similaire à *S2*, la solution retenue.

3. EXÉPRIENCES ET RÉSULTATS

Nous avons étudié les performances des systèmes acoustico-labiaux *S1*, *S2* et *S3*, comparativement au système acoustique de référence *S0* sur la tâche suivante : reconnaissance de séquences de 4 lettres prononcées de façon continue par un seul locuteur, extraites d'un corpus acoustico-labial. Réalisé à l'ICP de Grenoble ce corpus composé de 200 phrases représente l'enregistrement synchrone des informations acoustiques et des informations de suivi des lèvres. Deux tiers du corpus ont servi à l'apprentissage et un tiers aux tests.

Pour étudier les performances des systèmes en milieu bruité, nous avons ajouté au signal acoustique un « bruit de foule ». Le niveau de bruit correspond à un RSB de 10 dB, 0 dB et -10 dB.

Nous donnons (table 1) les résultats obtenus, en terme de taux « d'accurate » (c'est à dire le taux de lettres correctement reconnues moins le taux d'insertions), pour chacun des systèmes dans diverses conditions de bruit.

Table1 : Résultats des tests

RSB / Système	sans bruit	10 dB	0 dB	-10 dB
S0	90.8 %	85.5 %	62.3 %	-44.3 %[1]
S1	95.4 %	88.3 %	75.0 %	39.7 %
S2	96.1 %	89.0 %	77.4 %	44.3 %
S3	95.7 %	90.8 %	79.2 %	42.2 %

Les résultats obtenus confirment l'apport des informations labiales, surtout pour compenser la perte d'information dûe au bruit.

Une amélioration des performances des systèmes acoustico-labiaux *S1*, *S2* et *S3* par rapport au système acoustique de référence *S0* peut être constatée même dans un environnement non bruité. Cela pourrait s'expliquer par l'optimisation de l'intégration des deux modalités complémentaires : acoustique et visuelle, à l'aide de pondérations.

[1] Le taux « d'accurate » négatif arrive quand il y a plus d'insertions que des lettres correctement reconnues.

Le système *S1* résout le problème de différence de fréquence d'acquisition des paramètres acoustiques, par rapport aux paramètres labiaux (P3), en interpolant les paramètres labiaux. La pondération du flux labial par rapport au flux acoustique permet d'obtenir une amélioration dans le milieu bruité, sans pour autant introduire des erreurs de reconnaissance dans le cas d'un signal non bruité. À titre d'exemple, la valeur du coefficient de pondération du flux labial est de 0,5 à -10 dB et de 0,3 sans bruit.

La séparation des flux dans les systèmes *S2* et *S3* permet d'éviter ce même problème P3. Ces deux systèmes présentent de meilleures performances par la prise en compte du fait que le flux acoustique et le flux visuel sont la réalisation de phénomènes asynchrones (P2).

Contrairement à nos attentes, dans certains cas (sans bruit et à -10 dB), le système *S2* donne de meilleurs résultats que le système *S3*.

Cependant, la différence de performance n'étant pas toujours statistiquement significative, une confirmation de ces résultats reste à établir sur un corpus plus conséquent.

4. CONCLUSION

L'obtention de bonnes performances d'un système acoustico-labial de reconnaissance de la parole est conditionnée par une intégration optimale des deux modalités. Ce qui se traduit, dans la pratique, par la prise en compte de l'ensemble des problèmes exposés dans l'introduction.

De ce point de vue, la meilleure perspective est, nous le pensons, le modèle d'intégration du système *S3*, car il traite les phénomènes de rétention et d'anticipation labiale tout en utilisant les visèmes pour représenter le continuum labial. La raison pour laquelle ce système ne donne pas toujours les meilleures performances est probablement le fait que les visèmes que nous avons utilisés ont été définis dans un contexte perceptif et ne sont pas adaptés à la tâche de reconnaissance et au corpus. Nous travaillons actuellement sur la définition d'un ensemble de visèmes approprié.

Par ailleurs, une adaptation de la fonction de décision devrait améliorer les performances des systèmes *S2* et *S3*. Cette adaptation peut être faite dans un cadre probabiliste (par maximisation de la vraisemblance) ou au moyen des réseaux de neurones (par minimisation du taux d'erreur de reconnaissance).

5. BIBLIOGRAPHIE

Abry C., Lallouache M. T. (1991) *Audibility and Stability of Articulatory Movements. Deciphering two experiments on anticipatory rounding in French*, ICPS, n°1, 220-225

Adjouani A., Benoît C. (1995) *Audio-Visual Speech Recognition Compared Across Two Architectures*, Eurospeech, 1563-1566

C. Benoît C., Mohamadi T., Kandel S. (1994) *Effects of Phonetic Context on Audio-Visual Intelligibility of French*, Journal of Speech and Hearing Research, n°37, 1195-1203

Bregler C., Hild H., Manke S., Waibel A. (1993) *Improving Connected Letter Recognition by Lipreading*, ICASSP, n°1, 557-560

Cline A. (1974) *Scalar and Planar Valued Curve Fitting Using Splines Under Tension*, Communications of the ACM, v°17, n°4, 218-225

Ficher G. (1968) *Confusions Among Visually Perceived Consonants*, Journal of Speech & Hearing Disorders, n° 40, 481-492

Jacob B., André-Obrecht R., Parlangeau N.,. Sénac C. (1995) *Fusion des données acoustiques et articulatoires en reconnaissance automatique de la parole*, GRETSI, n°1, 365-368

Lallouache T. (1990) *Un poste visage-parole : Acquisition et traitement des contours labiaux*, JEP, Montréal

Robert-Ribes J. (1995) *Modèles d'intégration audiovisuelle de signaux linguistiques*, Thèse de doctorat, Institut National Polytechnique, Grenoble, France

Suaudeau N., André-Obretch R. (1993) *Sound Duration Modeling and Time variable Speaking Rate in a Speech Recognition System*, Eurospeech, 307-310

Summerfield Q. (1987) *Some Preliminaries to a Comprehensive Account of Audio Visual Speech Perception*, Dodd&Campbell editors, Hearing by Eye: the Psychology of Lipreading, 3-51

Jourlin P., El-Bèze M., Méloni H. (1995) Integrating Visual and Acoustic Information in a Speech Recognition System based on HMM, ICPhS, n° 4, 288-291

Jourlin P. (1995) *Produit de deux automates à transitions valuées, applications aux modèles de Markov cachés pour la reconnaissance de la parole bimodale et unimodale*, Rapport interne du LIA, 1995, 31-43

Un modèle Maître-Esclave pour la fusion des données acoustiques et articulatoires en Reconnaissance Automatique de la Parole

Bruno Jacob, Christine Sénac

Institut de Recherche en Informatique de Toulouse CNRS UMR 5505
Université Paul Sabatier

118, Route de Narbonne, 31062 Toulouse Cédex

ABSTRACT

In the project AMIBE (Applications Multimodales pour Interfaces et Bornes Evoluées), we study the natural visual and auditive bi-modality of the speech communication. The automatic speech recognition is performed by synchronizing the lip-reading and the acoustic pattern recognition based on Hidden Markov Models (HMM).

To merge acoustic and labial observations, we propose two alternatives :

- a classical HMM where the acoustic observations and the labial ones are assumed independent,
- a master-slave relation between two HMM, the articulatory HMM enslaves the labial one.

Automatic recognition experiments are performed on connected digit and spelled letter databases. We compare the two approaches and we show the lip-reading interest.

1. INTRODUCTION

Dans le cadre du projet Applications Multimodales pour Interfaces et Bornes Evoluées (projet AMIBE soutenu par les PRC Informatique 1993 - 1995), est étudiée la bimodalité naturelle auditive et visuelle de la communication orale. La reconnaissance automatique de la parole s'opère en synchronisant une "lecture labiale" avec un module de reconnaissance des formes acoustiques, par Modèles de Markov Cachés (MMC).

Pour fusionner les données acoustiques et articulatoires, plusieurs alternatives se présentent. Les informations peuvent être traitées sans discernement, par un MMC classique ; le vecteur d'observations est la concaténation des deux familles de paramètres labiaux et acoustiques ; ils sont considérés comme indépendants et l'utilisation de pondérations permet de réduire l'importance de l'une par rapport à l'autre (nous parlerons, dans la suite de cet article d'approche globale et de MMC global). Cette approche a été choisie par le LIUM avec un MMC dont les coefficients labiaux et acoustiques sont pondérés en fonction du bruit (Foucault, 95). Pour sa part, le LIA propose un nouveau MMC dont les unités sont des mots, gérant le décalage de façon interne grâce à un choix particulier de topologie (Jourlin, 95). Une autre alternative consiste à modéliser chaque famille de paramètres par un modèle de type MMC, et corréler les deux modèles par une dépendance entre les lois. Nous avons étudié plus particulièrement cette approche appelée par la suite approche maître-esclave ; nous nous sommes inspirés des travaux de Brugnara et De Mori qui ont appliqué cette liaison maître-esclave pour traiter la durée des sons (Brugnara, 92).

Au cours de cette présentation, le principe de l'approche maître-esclave est rappelé brièvement, et nous définissons un MMC équivalent moyennant certaines hypothèses simplificatrices appropriées à la reconnaissance de paramètres acoustiques et labiaux. Des résultats expérimentaux illustrent cette approche, dans les cas de la reconnaissance de suites de chiffres et lettres épelées. Une comparaison entre approche globale et approche maître-esclave est proposée.

2. PRINCIPE DU MODELE THEORIQUE

Le principe des modèles dits "maître-esclave" repose sur la modélisation d'une application non plus par un MMC unique, mais par deux MMC mis en parallèle et corrélés. L'idée générale est de parvenir à une adaptation dynamique des lois de probabilités d'un des modèles de Markov cachés, en fonction du contexte courant modélisé par l'autre MMC. Le contexte est une notion qui doit être prise au sens large, il peut s'agir d'un indice de voisement, de nasalisation,... d'un réel contexte phonétique, d'un indice suprasegmental comme la durée des sons... tandis que le MMC piloté est traditionnellement lié à des paramètres acoustiques.

Un modèle maître-esclave se compose de deux modèles : un modèle maître λ' qui est un MMC classique et un modèle λ'' qui

est un MMC dont les paramètres dépendent à tout instant de l'état dans lequel se trouve le modèle maître.

Un modèle maître-esclave est équivalent mathématiquement à un modèle classique de type MMC mais l'inconvénient de ce modèle réside dans son important nombre d'états et de lois. Ne pouvant raisonnablement implanter un tel modèle, nous avons réalisé une représentation simplifiée : nous réduisons le processus maître à un modèle ergodique dont les probabilités de transitions entre états ne sont pas réestimées.

Il s'en suit que nous pouvons créer un modèle simplifié dont le nombre d'états est le nombre d'états du modèle esclave, mais chaque transition du modèle esclave est dupliquée par le nombre d'états du modèle maître et chaque nouvelle transition est indexée par un état maître (Jacob, 95).

3. EXPERIMENTATIONS

Dans le cadre du projet AMIBE, nous disposons de deux types de signaux : le signal acoustique et le signal articulatoire synchronisés. Le signal acoustique est échantillonné à 16 kHz, tandis que pour le signal articulatoire (issu d'un traitement d'image (Lallouache, 91)), nous disposons d'un vecteur d'observations toutes les 20ms. Ce signal se compose de la largeur A ,de la hauteur B, et de la surface S intérolabiale .

3 . 1 . Pré traitement des données

Le signal acoustique est segmenté automatiquement (André-Obrecht, 88) et une analyse spectrale est faite sur chaque segment : 8 coefficients cepstraux (CC) sont obtenus après recalage du spectre selon l'échelle Mel. Leur sont adjoints l'énergie (E) et la dérivée de ces coefficients (Δ CC, Δ E). Les frontières issues de la segmentation statistique sont projetées sur les signaux articulatoires. Pour chaque segment projeté, est calculée une valeur moyenne de chaque paramètre labial ainsi que les dérivées correspondantes. Le vecteur d'observations est finalement composé de 18 coefficients de nature acoustique, de 6 coefficients articulatoires, auxquels est ajoutée la longueur du segment correspondant (T).

3 . 2 . Système de reconnaissance

Pour fusionner les données acoustiques et articulatoires, nous avons envisagé un modèle équivalent simplifié correspondant au modèle maître-esclave suivant :

— le modèle Maître est un modèle ergodique à 3 états modélisant les configurations des lèvres : ouvertes, fermées et semi-ouvertes.

— le modèle Esclave est un modèle gauche-droit. dont les unités acoustiques élémentaires sont des pseudo-diphones (André-Obrecht, 93).

Dans les deux séries d'expériences, le nombre de lettres , c'est-à-dire 4, est connu du système.

3.2.1. Données

Cette application de reconnaissance est monolocuteur et le système est évalué sur un corpus de lettres connectées : chacune des phrases est composée de 4 lettres épelées. L'apprentissage contient 158 phrases (soit 632 mots) et le test se compose de 48 phrases (soit 192 mots).

3.2.2. Résultats

3.2.2.1. Résultats en milieu calme

Afin de valider ce type d'approche, nous avons comparé systématiquement les taux de reconnaissance à ceux obtenus à l'aide d'un modèle de Markov Caché Global construit de manière classique en utilisant aussi le pseudo-diphone comme unité élémentaire. Un vecteur d'observations est traité globalement, à raison d'une loi gaussienne par transition (matrice de covariance diagonale).

Le MMC Global est appris initialement avec 8 coefficients cepstraux, l'énergie et la durée. Le taux de reconnaissance avec ces seuls paramètres acoustiques est de 89,6%. Nous avons ajouté successivement les paramètres labiaux et leurs dérivées. La même expérience a été répétée en initialisant le modèle global avec 8 coefficients cepstraux, les dérivées des quatres premiers coefficients, l'énergie ainsi que sa dérivée, et la durée du segment. Le meilleur taux de reconnaissance, à savoir **91,6 %** (taux mots) est obtenu en utilisant la hauteur et la largeur des lèvres (figure 1.a). L'introduction de la surface des lèvres n'apporte pas d'information pertinente car elle est fortement corrélée aux paramètres A et B (Benoît, 91). Les dérivées des coefficients labiaux dégradent le taux de reconnaissance : une des causes principales peut être le manque de synchronisation entre les informations labiales et acoustiques, ou le manque de données d'apprentissage.

Le même protocole d'expérimentation est réalisé pour tester l'approche Maître-Esclave. Le coefficient labial A fait partie des paramètres initiaux. Le meilleur taux de reconnaissance est obtenu par le modèle ayant pour paramètres 8CC, E, T, A et B, à savoir **91,7** % en terme de mots correctement reconnus(figure 1.b). Lorsque le nombre de paramètres augmente, les performances décroissent, la cause est très certainement liée au relativement faible ensemble de données d'apprentis-

sage par rapport au nombre de paramètres à apprendre.

3.2.2.2. Résultats en milieu bruité à 15 dB

Afin de réaliser cette étude, nous reprenons les expériences du corpus des lettres en bruitant artificiellement à 15 dB les fichiers contenant le signal. Le bruit utilisé est de nature "cocktail party".

Nous n'avons pas utilisé la segmentation automatique, afin de ne pas ajouter les problèmes dus aux erreurs de segmentation à notre problème initial. Les observations acoustiques sont donc centisecondes et les modèles acoustiques des unités pseudo-diphones liés à un pré-traitement segmental sont remplacés par des modèles acoustiques plus adaptés à un découpage centiseconde.

Etant donné que la durée du segment est maintenant constante, le paramètre T ne fait plus partie des paramètres du modèle Global et du modèle Maître-Esclave.

Dans le modèle Global, les paramètres de base sont les coefficients cepstraux auxquels nous ajoutons progressivement des paramètres labiaux et/ou acoustiques. Les résultats sont montrés par la figure 1.c. Le meilleur taux de reconnaissance que nous obtenons est de **78,7%** (soit 41 lettres sur 192 non correctement reconnues). Ce modèle n'utilise que les paramètres de base et les paramètres labiaux A et B. L'introduction des dérivées qu'elles soient de paramètres acoustiques ou labiaux n'apporte rien aux performances du système. Nous constatons que l'apport de l'énergie E dégrade les taux de reconnaissance en milieu bruité.

Dans le modèle Maître-Esclave, le paramètre labial A et les quatres premiers coefficients cepstraux constituent les paramètres de base. On ajoute progressivement dans les modèles Maître et Esclave les paramètres labiaux et acoustiques. Etant données les dégradations observées dans le modèle Global, l'énergie E du signal n'a pas été retenue dans le choix des paramètres. Les résultats sont donnés dans la figure 1.d. Le meilleur taux de reconnaissance obtenu est de **77,6%** (c'est-à-dire 43 lettres incorrectement reconnues sur 192). Comme dans le modèle Global, les paramètres maximisant le taux de reconnaissance sont les 8 coefficients cepstraux ainsi que les paramètres labiaux A et B.

Les expériences réalisées en milieu bruité ont indiqué que les résultats obtenus avec les deux modèles Global et Maître-Esclave sont là encore comparables.

Cependant, il faut se rappeler de l'augmentation du nombre de paramètres de ces modèles dûe à une plus grande complexité des modèles acoustiques centisecondes par rapport aux modèles acoustiques segmentaux. L'ensemble d'apprentissage restant le même, ces modèles n'ont bénéficié que d'un apprentissage moindre. Si l'on tient compte du grand intervalle de confiance et de l'explosion combinatoire du nombre de paramètres du modèle Maître-Esclave par rapport au modèle Global, les résultats obtenus sont alors comparables.

4. CONCLUSION

Nous avons présenté deux approches probabilistes pour traiter la fusion de données acoustiques et articulatoires dans un but de reconnaissance. L'approche classique consiste à supposer les informations issues des deux canaux indépendantes tandis que l'approche maître-esclave exploite une certaine corrélation par l'intermédiaire de liens entre les lois d'observation.

L'approche du LIUM permet d'obtenir des taux de précision ("accurate") de 96% en ambiance calme et de 91% à 10dB. Le LIA réalise des scores de reconnaissance en termes de lettres de 90% en ambiance calme. Les deux approches *modèle global* et *modèle maître-esclave* donnent des résultats très comparables dans le cadre de la reconnaissance mono locuteur de lettres épelées (92 % de taux de reconnaissance en mots dans une ambiance calme et 78% dans le bruit à 10dB). L'avantage de la deuxième méthode est liée à une meilleure compréhension du phénomène labial et offre des perspectives intéressantes :

— Au niveau maître, nous augmenterons le nombre d'états de manière à se rapprocher des études statistiques qui ont montré l'émergence de visèmes (Benoît, 91)

— Le modèle maître-esclave est, dans son actuelle implémentation, fort simplifié et certaines hypothèses sont trop fortes : le passage d'un état *ouvert* à celui de *fermé* ne se réalise pas de manière instantanée! En fonction du volume croissant de l'ensemble d'apprentissage qui nous sera ultérieurement fourni, nous complexifierons le modèle simplifié pour tendre vers le modèle exact et tester ses réelles possibilités.

—L'étude d'une désynchronisation entre le labial et l'acoustique est plus abordable par cette approche.

L'utilisation des paramètres labiaux a pour but de rendre plus robuste la reconnaissance automatique de parole en milieu bruité ; nous avons montré que cette information ne dégradait absolument pas les performances des systèmes actuels déjà très performants. Nous étudions actuellement des rapports signal sur

bruit plus faibles.

REFERENCES

R. André-Obrecht (1988) : A new statistical approach for the automatic segmentation of continuous speech signals, *IEEE Trans. on Acoustics, Speech, Signal Processing*, vol. 36, n°1, janvier 1988.

R. André-Obrecht (1993) : Segmentation et parole? *Habilitation à diriger des recherches, IRISA*, Rennes, juin 1993.

C. Benoît (1991), C. Abry, L.J. Boë : The effect of context on labiality in french. *Eurospeech*, Gènes, 1991

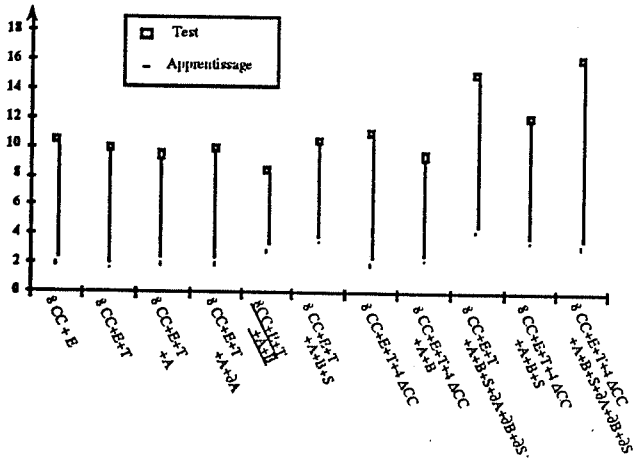
F. Brugnara (1992), R. De Mori, D. Guiliani, M. Omologo : A family of Parallel Hidden Markov Models, *ICASSP 92*, San Francisco, 1992.

A. Foucault (1995) : Système acoustico-labial de reconnaissance de la parole, *GFCP-SFA Journées Jeunes Chercheurs*, ENST Paris, 1995.

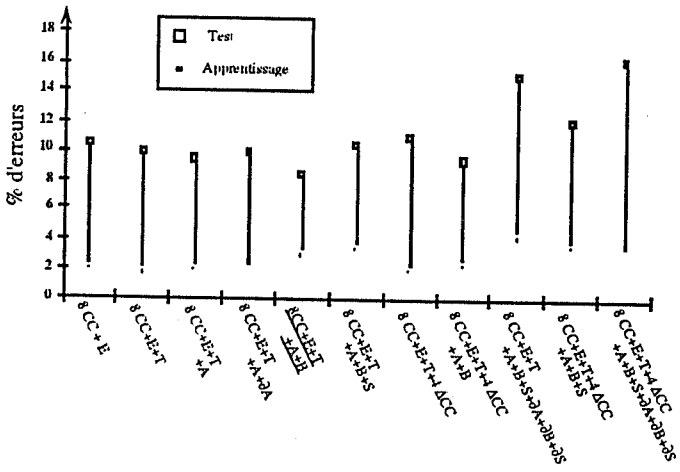
P. Jourlin (1995) : Automatic bimodal speech recognition, *ICPhS-95*, Stockholm, 1995.

B. Jacob (1995) : Un outil informatique de gestion des Modèles de Markov cachés : expérimentation en reconnaissance automatique de la parole, *Thèse de 3°cycle*, Toulouse III, 1995.

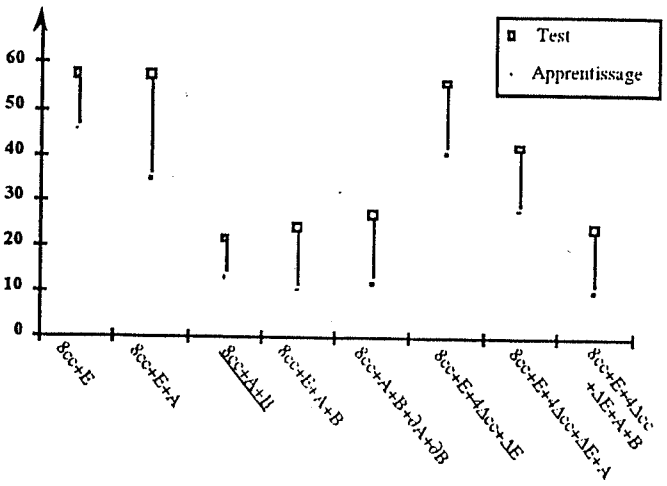
T. Lallouache (1991) : Un poste "visage parole" couleur. Acquisition et traitement automatique des contours de lèvres, *Thèse de doctorat de l'Institut National Polytechnique de Grenoble*, 1991.



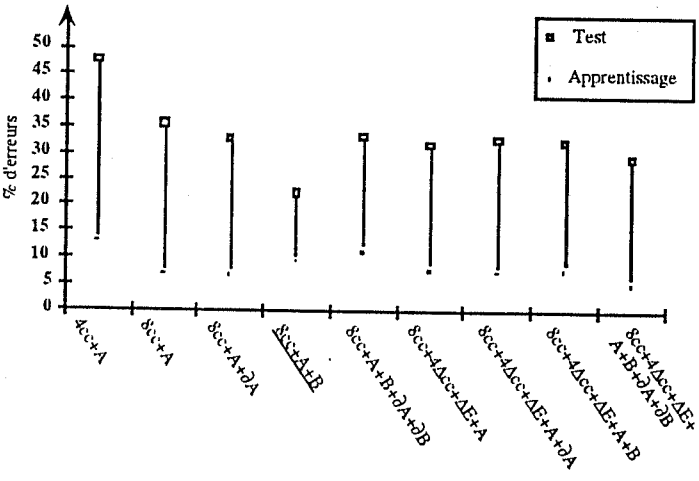
a) MMC global



b) MMC Maître/Esclave



c) MMC Global à 15 dB



d) MMC Maître-Esclave à 15 dB

Figure 1 : Taux d'erreurs en termes de lettres

JEP 96

SYNTHÈSE

AVIGNON 10-14 JUIN 1996

EVALUATION D'UN MODÈLE DE SOURCE DE FRICTION POUR LA SYNTHÈSE ARTICULATOIRE DES CONSONNES FRICATIVES

Khaled Mawass, Pierre Badin, Christophe Vescovi & Denis Beautemps

Institut de la Communication Parlée, UPRESA CNRS 5009, INPG – Université Stendhal
46, Av. Félix Viallet, F-38031 Grenoble Cedex 01 – Fax: 76 57 47 10 – e-mail: mawass@icp.grenet.fr

ABSTRACT

Fricative consonants involve the excitation of the vocal tract by noise sources: the air flow through a constriction in the tract causes a turbulent jet that generates such noise sources, when impinging on an obstacle or following a wall. The aim of the present study was to implement in an articulatory synthesiser a noise source model elaborated from data on fricatives in vocalic context (Badin et al., 1995b) and to assess it through the articulatory synthesis of fricative consonants. Starting from a corpus of midsagittal profiles obtained by cineradiography in synchrony with from lip video pictures, articulatory trajectories were generated, and used, in conjunction with aerodynamic data acquired on the same corpus and for the same subject, as commands to the synthesiser. This resulted in fairly high quality synthesis of fricative consonants, as will be demonstrated at the conference.

1. INTRODUCTION

Les consonnes fricatives font intervenir une excitation du conduit vocal par des sources de bruit: l'écoulement de l'air à travers une constriction dans le conduit vocal provoque un jet qui peut générer, lorsqu'il heurte un obstacle ou longe une paroi (Shadle, 1991), ces sources de bruit de friction, dont les caractéristiques dépendent de la géométrie et de l'état aérodynamique du conduit.

Un premier modèle de source de bruit a été développé à partir de données sur des fricatives soutenues (Badin, 1994). Plus récemment, ce modèle a été étendu à des fricatives en contexte vocalique (Badin et al., 1995b).

L'objectif de cette étude était d'implémenter ce dernier modèle dans un synthétiseur articulatoire, et de le tester dans le cadre de la synthèse de consonnes fricatives.

1. LE MODÈLE DE SOURCE DE FRICTION

A l'heure actuelle, la prédiction des bruits générés par un écoulement turbulent reste un problème extrêmement complexe, seules des descriptions qualitatives étant disponibles au niveau théorique. Ceci est en particulier dû aux phénomènes très fortement non-linéaires et instationaires mis en oeuvre ainsi qu'à la grande importance du couplage fluide-structure.

Dans la présente étude, nous faisons l'hypothèse que la source de friction est indépendante de la configuration du conduit vocal, et qu'elle est localisée au niveau des incisives pour les fricatives du français. La modélisation développée par Badin et al. (1995b) consiste à prédire le niveau global SPL et la pente spectrale moyenne TLT de la source de bruit en fonction de la chute de pression ΔP et de l'aire Ac à la constriction. Le modèle a donc été établi par analyse statistique des données mesurées pour un corpus de séquence [pVFV] prononcées par un locuteur français (PB). Les fricatives F = [f s ʃ] sont combinées symétriquement avec les voyelles cardinales V = [a i u].

1.1. Obtention des données

La chute de pression ΔP à la constriction est mesurée par un transducteur de pression connecté à un étroit tube de polyéthylène dont l'extrémité débouche en amont de la constriction.

Le débit aux lèvres U est obtenu à l'aide d'un masque pneumotachométrique de Rothenberg. Un équivalent aérodynamique de l'aire minimum de constriction Ac est obtenu grâce à l'équation dite de l'orifice, dérivée de l'équation de Bernoulli (Scully, 1986):

$$Ac = \frac{U}{\sqrt{2 \cdot (\Delta P / \rho)}},$$

où $\rho = 1.14 \cdot 10^{-3}$ g/cm^3 est la densité de l'air.

En première approximation, et pour une gamme de fréquence inférieure à 5-6 kHz, la source de bruit peut être caractérisée par son niveau SPL global et par la pentre moyenne du spectre TLT exprimée en dB/Octave (Badin et al., 1994).

Pour déterminer le spectre de la source, les techniques traditionnelles de filtrage inverse ne peuvent pas être utilisées, car seule la partie pôle de la fonction de transfert du conduit vocal est connue, mais pas la partie zéro. Une technique de filtrage inverse simplifié a donc été développée par Badin (1994), pour déterminer

les variations du spectre de source en fonction de ΔPc et de Ac.

1.2. Modélisation de la source de friction

Le seul modèle de source de friction proposé dans la littérature est celui de Stevens (1971), qui calcule l'amplitude de la source SPL (niveau spectral de puissance) en fonction des paramètres aérodynamiques ΔP et Ac. Badin et al. (1994) ont montré plus récemment que la pente spectrale globale TLT (définie en dB/Octave) varie aussi en fonction de l'intensité du son.

Le modèle plus complet développé par Badin et al. (1995b) à partir de fricatives en contexte vocalique prédit la variation de SPL et de la pente spectrale globale de la source, en fonction de ΔP et Ac:

$$SPL = K_1 + p \cdot log_{10}(\Delta P) + q \cdot log_{10}(Ac)$$

$$PNT = K_2 + r \cdot log_{10}(\Delta P) + s \cdot log_{10}(Ac)$$

Ce modèle de variation de la source a été évalué en comparant les spectres tiers d'octave prédits avec ceux mesurés. L'erreur absolue sur SPL varie entre 2 et 3 dB; l'erreur sur TLT peut atteindre 0.5 dB/Octave.

Le modèle final de source de bruit est obtenu en appliquant le modèle de variation défini ci-dessus à une source de base constituée d'un bruit blanc gaussien filtré passe-bas à 8 kHz.

Les sources de bruit générées par turbulence peuvent être, de manière générale, distribuées dans le conduit vocal. Cependant, en première approximation, pour les fricatives du français, on peut considérer une source ponctuelle localisée au niveau des incisives supérieures (Shadle, 1991). Cette source est insérée comme une source de pression en série dans le conduit vocal.

2. LE SYNTHÉTISEUR: MODÈLE AÉRODYNAMIQUE ET SOURCE DE FRICTION

2.1. Description générale

Un modèle de simulation temporelle du conduit vocal a été mis en œuvre à l'ICP. Il tente de représenter les phénomènes acoustiques, aérodynamiques et mécaniques intervenant dans la production de la parole. La simulation s'appuie sur les équations caractérisant la propagation du son dans le conduit vocal.

2.2. Modèle de voisement

Le modèle de voisement est basé sur un modèle simplifié de cordes vocales (Vescovi et Castelli, 1995). Chaque corde vocale, de longueur totale Lg, est divisée en une partie vibrante et une partie non vibrante. Les gestes d'abduction-adduction de la glotte sont ainsi définis par deux paramètres de contrôle : (1) le pourcentage de la longueur de la partie vibrante par rapport à Lg, (2) la distance au repos entre les parties vibrantes. Une relation linéaire entre ces deux paramètres a été imposée de façon à simplifier le contrôle du modèle.

2.3. Modèle aérodynamique simplifié

Le modèle de source de friction est contrôlé par deux paramètres aérodynamiques, ΔP et Ac. Une modélisation aérodynamique simplifiée, valable pour les basses fréquences réduit le conduit vocal à deux constrictions: la glotte et la constriction orale. Les équations de Bernoulli et de Poiseuille permettent alors d'exprimer la chute de pression ΔP à la constriction orale, considérée comme une fente parallélépipédique de largeur l_c, de hauteur $h_c = l_c/5$ et de longueur d_c dans le sens du débit par :

$$\Delta P = 0.5 \cdot \rho \cdot \frac{U^2}{Ac^2} + 12 \cdot \mu \cdot \frac{d_c \cdot l_c^2 \cdot U}{Ac^3}$$

où $\mu = 1.84 \cdot 10^{-4}$ dynes·s/cm² est la constante de viscosité dynamique, et la chute de pression ΔP_g à la glotte par une équation similaire (largeur l_g, longueur d_g, Ag est la composante basse fréquence, obtenue par filtrage passe-bas à 80 Hz de l'aire glottique Ag_{M2M} calculée dans le modèle à deux masses). La pression subglottique Ps est alors obtenue comme $P_s = \Delta P_g + \Delta P$.

2.4. Insertion de la source de bruit

Nous avons vérifié, sur quelques cas simples, que la transconductance entre débit acoustique aux lèvres et source de pression insérée en série au niveau des incisives, calculée comme la réponse impulsionnelle du synthétiseur lorsque la glotte est fermée, est semblable à celle calculée par le modèle fréquentiel VCTR (Badin, 1989).

3. LE CORPUS DE TEST

Un corpus contenant des combinaisons [pVCV] avec V= [a i u y] et C = [v z ʒ] a été utilisé pour resynthétiser des consonnes fricatives et ainsi tester le modèle de source, ainsi que son contrôle. Six contextes vocaliques ont été choisis: aCa, aCi, aCu, iCi, iCu et iCy. Pour ce corpus, nous avons enregistré avec le même sujet de référence PB des données articulatoires et aérodynamiques.

3.1. Données articulatoires

Les données articulatoires consistent en un ensemble d'environ 1200 contours sagittaux, tracés manuellement à partir d'un film cinéradiographique (à 50 images/s), synchronisés avec des images de face des lèvres du sujet

(Badin, 1995a). La fonction d'aire, calculée à partir de la fonction sagittale par un modèle de passage développé par Beautemps et al. (accepté), permet de calculer finalement le signal acoustique ou les formants.

3.2. Données aérodynamiques

Les données aérodynamiques brutes sont constituées par le débit aux lèvres U, mesuré à l'aide d'un masque pneumotachographique de Rothenberg, et par la chute de pression ΔP à travers la constriction.

4. STRATÉGIE DE DÉTERMINATION DES PARAMÈTRES DE COMMANDE

Pour effectuer la synthèse d'une séquence VFV, il est nécessaire de déterminer l'évolution temporelle de la fonction d'aire du conduit vocal (y compris l'aire minimale de constriction Ac) et de la pression sublottique Ps, et de l'aire au repos de la glotte Ag_0. Pour les consonnes fricatives dans le corpus étudié, Ac varie entre 0.05 et 0.18 cm^2. L'estimation de Ac à partir des données cinéradiographiques est relativement mauvaise, car le bruit absolu sur la fonction sagittale est important (de l'ordre de 1.5 mm), dû à l'ensemble de la chaîne de détection manuelle des contours à partir des images radiographiques, la Figure 1 permet de comparer les aires minimales obtenues par les deux méthodes.

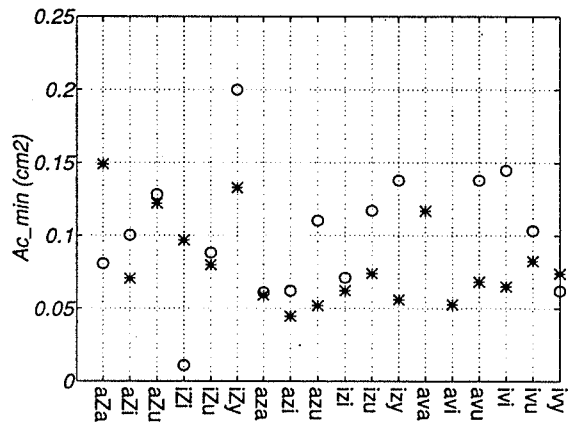


Figure 1 – Comparaison entre le Ac "cinéradiographique" (o) et le Ac "aérodynamique" (*) pour les fricatives voisées.

Plus précisémment, la distance sagittale dans la région de constriction des fricatives étant de l'ordre 2 mm, l'erreur relative peut atteindre 75%. Pour résoudre ce problème, nous avons remplacé, pour les petites valeurs de Ac (en dessous de 0.2 cm^2), l'estimation de Ac basée sur les données cinéradiographiques par l'estimation basée sur les données aérodynamiques pour le même corpus et le même sujet PB. Après avoir aligné temporellement les signaux à l'aide du paramètre d'énergie filtrée passe-bas à 10 Hz, une fenêtre d'émergence a été utilisée pour forcer la trajectoire de Ac d'origine cinéradiographique à suivre la trajectoire d'origine aérodynamique au centre de la consonne, tout en évitant toute discontinuité. La fenêtre choisie est une gaussienne centrée à l'instant où Ac est minimum, et dont la variance dépend de la durée de la fricative. Un exemple de résultat se trouve à la Figure 2.

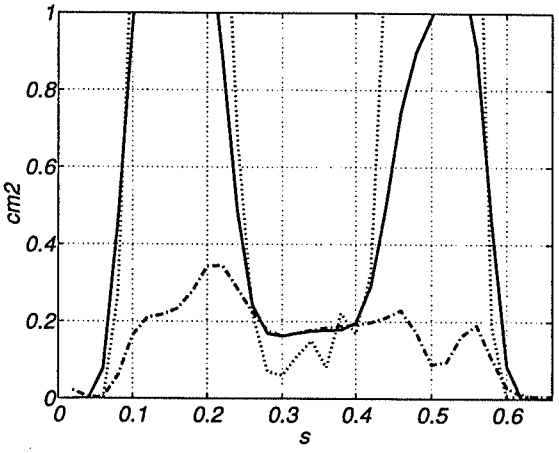


Figure 2 – Exemple de résultat de fusion (—) entre Ac cinéradiographique (...) et aérodynamique (–.–.) pour [paSa].

En supposant que la glotte est suffisamment ouverte lors d'un [p], la pression sousglottique Ps est égale à la pression intraorale ΔP durant le [p]. Nous avons donc estimé Ps comme la moyenne de ΔP entre le [p] précédent et le [p] suivant la séquence VFV (Scully, 1986). Grâce aux équations aérodynamiques simplifiées ci-dessus, la composante basse fréquence de l'aire glottique Ag peut être facilement estimée à partir de Ps, ΔP et U. L'erreur d'estimation de Ag est liée directement à l'approximation de Ps. Nous avons vérifié que, dans le cas des fricatives non voisées, une variation importante de la valeur estimée de Ps (entre 10 et 13 cmH_2O) maintient un Ag supérieur à 0.2 cm^2, ce qui assure le dévoisement, et ne joue pas de rôle dans la génération du bruit de frication. Le cas des fricatives voisées, où la valeur absolue de Ag joue un rôle important pour le voisement, est plus délicat, et nécessite d'affiner l'estimation de Ag de manière plus indirecte, en vérifiant en particulier le degré de voisement sur le signal rayonné lui-même.

5. RÉSULTATS ET PERSPECTIVES

En utilisant la stratégie décrite dans la section précédente, et en faisant l'hypothèse que l'articulation des fricatives non voisées [f s ʃ] est très proche de celle des fricatives voisées correspondantes [v z ʒ], sauf éventuellement au niveau de l'aire de constriction, nous avons déterminé de manière automatique les trajectoires temporelles de la fonction d'aire, de Ac, Ps, et Ag pour les contextes vocaliques

aFa, aFi, aFu, iFi, iFu et iFy. La Figure 3 montre les trajectoires de ces paramètres pour les séquences [paʃa] et [paʒa].

L'examen des résultats appelle un certain nombre de remarques.

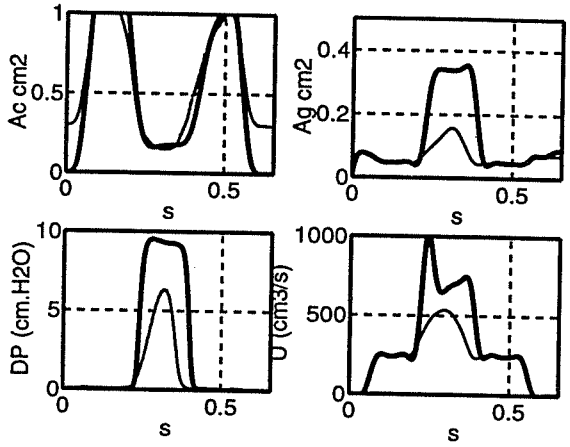


Figure 3 – Exemple de trajectoires des paramètres Ac, Ag, ΔP et U pour [paʃa] (épais) et [paʒa] (fin).

Un jugement informel d'écoute a montré que la plupart des fricatives sont de très bonne qualité. Quelques problèmes d'articulation étaient liés de manière claire à des imperfections au niveau des transitions de la fonction d'aire.

La valeur de Ag atteinte au centre de la fricative n'est pas critique pour la source de bruit pourvu qu'elle soit au dessus de 0.2-0.3 cm² (ceci est confirmé par des simulations).

La relation entre Ag et les trois paramètres de contrôle des gestes d'abduction-adduction de la glotte est importante pour que les transitions du signal de débit glottique simulé soient suffisamment semblables à celles observées sur le signal naturel (ces différences peuvent être analysées en comparant les signaux synthétiques et naturels de pression rayonnée aux lèvres).

Pour quelques fricatives voisées, la valeur de Ag dans la partie fricative était surestimée, dû à une sous-estimation de la pression sousglottique Ps, ce qui conduisait à un dévoisement. Une diminution de Ag a permis de remonter le degré de voisement à un niveau correct.

En conclusion, la présente étude montre la faisabilité de la synthèse articulatoire des fricatives, et la validité du modèle de source de bruit développé par Badin et al. (1995b), du moins en ce qui concerne les fréquences jusque vers 5-7 kHz.

L'estimation de Ag n'est pas très précise, et il serait souhaitable soit d'avoir un accès direct soit à Ps, soit à Ag. Par ailleurs, il est nécessaire d'étendre la gamme de fréquence à des fréquences de l'ordre de 10-12 kHz: cela nécessite un modèle acoustique plus sophistiqué, qui prenne en compte les modes de propagation acoustique d'ordre supérieur. Il faut enfin noter que le dispositif expérimental nécessaire pour obtenir les données est très lourd: l'avenir nous semble reposer sur les techniques d'inversion qui permettent d'inférer les gestes articulatoires à partir du signal de parole (Abry et al., 1994).

6. REMERCIEMENTS

Ce travail a été partiellement financé par le projet européen ESPRIT/BR *Speech Maps*.

7. RÉFÉRENCES

Abry, C., Badin, P., & Scully, C. (1994) Sound-to-gesture inversion in speech: The Speech Maps approach. In *Advanced speech applications* (Varghese K. et al., Eds), pp. 182-196. Springer.

Badin P. (1989) Acoustics of voiceless fricatives: production theory and data. STL-QPSR 3/1989, 33-55.

Badin, P., Gabioud, B., Beautemps, D., Lallouache, T.M., Bailly, G., Maeda, S., Zerling, J.P., & Brock, G. (1995a) Cineradiography of VCV sequences: Articulatory-acoustic data for a speech production model. *ICA*, Trondheim, Norway, June 1995, Vol. IV, 349-352.

Badin, P., Mawass, K., & Castelli, E. (1995b). A model of frication noise source based on data from fricative consonants in vowel context. *XIIIth ICPhS*, Vol. 2, 202-205.

Badin, P., Shadle, C.H., Pham Thi Ngoc, Y., Carter, J.N., Chiu, W., Scully, C., & Stromberg, K. (1994) Frication and aspiration noise sources: contribution of experimental data to articulatory synthesis. ICSLP, Yokohama, Japan, Vol.1, paper S06-4, 163-166.

Beautemps, D., Galvan, A., Badin P., Bailly, G., & Laboissière, R. (accepté) Evaluation of an articulatory-acoustic model based on a reference subject. 4th Speech Production Seminar, Autrans, France, 1996.

Shadle, C.H. (1991) The effect of geometry on source mechanisms of fricative consonants. *Journal of Phonetics* 19, 409-424.

Stromberg, K., Scully, C., Badin, P., & Shadle, C.H. (1994) Aerodynamic patterns as indicators of articulation and acoustic sources for fricatives produced by different speakers. *Proc IOA*, Vol.16-5, pp 325-333.

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

Stevens, K.N. (1971) Airflow and Turbulence Noise for Fricative and Stop Consonants: Static Considerations. J. Acoust. Soc. Am. 50, 1180-1192.

Vescovi, C., & Castelli, E. (1995) A robotic approach to the inversion of a two-mass model. In Shadle (Ed.)*From speech signal to acoustic sources*, PPR N°3, European ESPRIT/BR *Speech Maps* project. Vol. II.

LES LIAISONS ET LA SYNTHESE VOCALE

Philippe BOULA de MAREÜIL

LIMSI-CNRS - BP 133 - 91403 Orsay Cedex
Tél.: 69 85 81 02 - Fax: 69 85 80 88 - e-mail: mareuil@limsi.fr

ABSTRACT

In this paper, the problem of the liaison in French is approached, in a general frame and in the application of text-to-speech systems. The phonological particularities of this phenomenon are studied, as well as when the liaison is made, especially within expressions. A few results provided by 5 graphemic-phonemic converters are then provided.

1. INTRODUCTION

Des consonnes muettes en français contemporain, en position faible ou dans les mots isolés, peuvent apparaître lorsque le mot qui suit immédiatement commence par une voyelle graphique ou un *h* muet, évitant ainsi un hiatus. C'est le fait de liaison, survivance d'une époque où toutes les consonnes finales se faisaient entendre.

Dans cet article, nous confrontons ce phénomène à ceux de l'enchaînement et de l'élision. Nous décrivons son fonctionnement : comment et quand fait-on la liaison? On est là en terrain délicat, si bien qu'il n'y a pas consensus pour trancher cette question. Nous interpréterons ensuite un corpus fermé pour comparer quantitativement 5 synthèses vocales, entre elles et avec les résultats données par des sujets parlants.

2. LIAISON, ENCHAINEMENT ET ÉLISION

Par rapport à la liaison, l'enchaînement intervient dans les mêmes conditions, mais concerne les consonnes finales normalement prononcées (Fouché, 1969). On peut même entendre des liaisons sans enchaînement, particulièrement dans les débats politiques (Encrevé, 1988).

Il faut également distinguer la liaison et l'élision qui, au lieu d'ajouter une consonne, supprime une voyelle ; celle-ci peut être remplacée dans l'écriture par une apostrophe.

Il convient néanmoins de noter que dans les 3 cas, la consonne qui termine le premier élément appartient généralement à la syllabe initiale du mot suivant, même si un blanc, une apostrophe ou un trait d'union la sépare, à l'écrit, de la voyelle sur laquelle elle s'appuie. Cela peut rendre difficile la reconnaissance des frontières de mot et donc l'accès au lexique (Wauquier-Gravelines, 1994).

Ces définitions posées, il serait intéressant de voir ce qu'il en est perceptivement : à l'oreille, un Français fait-il la différence entre

(a) "Claude n'a pas de petit ami" ;

(b) "Claude n'a pas de petite amie" ;

(c) "Claude n'a pas de petit tamis"?

3. COMMENT FAIT-ON LA LIAISON?

Dans les liaisons proprement dites, certaines modifications phoniques peuvent avoir lieu. Si les consonnes *-z*, *-p* et *-t* s'articulent conformément à l'orthographe, il n'en va pas systématiquement ainsi. Certaines consonnes sourdes (fricatives) deviennent sonores :

's' et 'x' -> /z/ (ex. *les/aux enfants*),

'f' -> /v/ dans *(vingt-)neuf ans* et *(dix-)neuf heues* ;

et inversement (pour les occlusives) :

'd' -> /t/ (ex. *un grand homme*).

Il y a lieu de souligner enfin que les groupes figés se disent souvent au pluriel comme au singulier (ex. *des guets-apens*) - on s'abstient donc aussi de dire avec liaison *salles à manger*, *arcs-en-ciel*, *machines à coudre*, etc. Il existe quelques exceptions à cette règle, comme *(arrière-)petits-enfants* ou *grands-oncles*. Nous avons néanmoins relevé dans BDLEX (Pérennou & Calmès, 1987) 20 formes en *-s* trait d'union voyelle - apparaissant lors du passage au pluriel de mots composés, sans présenter de liaison. Elles ont été consignées dans un dictionnaire d'exceptions.

Dans l'ensemble, la liaison n'a pas d'effet sur la voyelle précédente, mais, avec la resyllabation, elle peut entraîner la dénasalisation de /ɛ̃/ et /ɔ̃/ : on prononce /divin/ dans *le divin Enfant*. /ɛ̃/ se change naturellement en /ɛn/ dans tous les adjectifs en *-ain* (ex. *certain*, *vilain*, *vain*, *prochain*, *lointain*, *soudain*), en *-ein* (ex. *plein*), en *-ien* (ex. *ancien*) et en *-yen* (ex. *moyen*). Les adjectifs finissant ainsi sont au nombre de 83 dans BDLEX. Mais nous venons de fournir ceux qui sont susceptibles d'être antéposés. Quant à *rien* (ex. *rien à faire*) et *bien*, ils gardent leur nasalité (ex. *bien habillé*).

/ɔ̃/ perd sa nasalité dans *bon* (ex. *bon ami*) ainsi que dans des mots comme *non-intervention, non-éxécution.*

Quant aux voyelles orales, il n'y a de remarque à faire que pour les masculins singuliers finissant en *-er*. Alors qu'on prononce un /e/ dans *léger* et *dernier* en finale absolue, le /E/ s'ouvre sous l'influence du /R/ (ex. *dernier homme*).

4. QUAND FAIT-ON LA LIAISON?

Dans le français de la conversation et de la lecture courante, la liaison est un problème épineux, en pleine évolution, fluctuant avec le style : d'autant plus rare que le langage est ordinaire, familier ou populaire, sa fréquence augmente à mesure que l'on s'approche du discours soutenu, affecté ou littéraire - peut-être par souci d'intelligibilité - (Donohue-Gaudet, 1969). D'où la distinction traditionnelle entre liaisons obligatoires (ou invariantes), facultatives (ou variables) et interdites (ou erratiques). Mais il subsiste beaucoup d'hésitations en synchronie, la répartition en ces différentes catégories dépendant du milieu socio-culturel et du niveau de langue : des liaisons possibles peuvent être indispensables dans une déclamation sur la scène tragique, dans une diction soignée de conférence ou dans la récitation de la poésie classique et romantique, alors qu'elles sont choquantes dans un usage relâché, ou semblent artificielles, snob, voire ridicules, surtout s'il y en a trop (Eggs & Mordellet, 1990).

La tendance actuelle retient cependant les liaisons qui ont un rôle grammatical. La liaison - que nous noterons par le symbole " _ " tandis que son absence sera indiquée par une barre verticale "|" - peut en effet opposer un pluriel à un singulier : comparez *prix_élevés* et *prix|élevé*, *la vente d'armes_anglaises* et *la vente d'armes|anglaise*, etc. Son absence peut aussi servir à distinguer un substantif d'un adjectif (ex. bien connu *un savant|aveugle* vs *un savant_aveugle*), ou encore des mots avec ou sans *h* aspiré : *des héros - des zéros* (alors que le *h* est muet dans *héroïne*), *les hauteurs - les auteurs*, *ces hêtres - ces êtres*, *en haut - en eau.*

Devant la complexité des liaisons facultatives (peu de spécialistes y ont d'ailleurs accordé leur attention, dans leurs recherches), notre tâche se bornera à l'étude des liaisons qualifiées d'obligatoires et d'interdites. Nous ne nous attarderons pas à examiner la place qu'occupent le registre de langue, le débit et la situation, dans cette division : nous comptons uniquement ce qui est jugé acceptable et ce qui ne l'est pas dans l'application que nous envisageons, à savoir la synthèse vocale.

D'ailleurs, "facultatif" ne signifie pas nécessairement "aléatoire", et à moins d'introduire des variations, deux démarches seulement - déterministes - sont possibles en synthèse de la parole : elles consistent à réaliser toutes les liaisons qui ne sont pas interdites, ou bien ne réaliser que les liaisons obligatoires. Nous opterons pour la 2e solution - "minimaliste" - mais que Divay et Guyomard (1977) ont justifiée, ayant essayé les deux.

On est, en effet, loin de lier dès que le contexte consonne-voyelle s'y prête. Il faut affiner cette hypothèse : introduire une règle supplémentaire, stipulant un rapport grammatical étroit entre les deux termes en contact - lequel s'exprime en général par l'appartenance à un même groupe rythmique. Ce lien est établi dans la construction déterminant + déterminé (Léon, 1976). En général sinon, on évite la liaison entre syntagmes nominal et verbal - fonction démarcative participant, de la même manière que la prosodie, au découpage du flux sonore et à la hiérarchisation du message.

Les liaisons sont gouvernées par la longueur des mots, mettant à jour des propriétés morphologiques ; elles révèlent en outre des relations syntaxiques et permettent en partie d'évaluer le degré de cohésion entre les mots - comme les facteurs prosodiques, avec lesquels elles sont corrélées (ceci expliquant cela) - (Lucci, 1983).

On respecte la liaison dans la plupart des locutions toutes faites, qui constituent des sortes de mots composés : comparez par exemple *avoir un pied à terre* et *avoir un pied_à-terre*. Il résulte de là que certains syntagmes, à l'intérieur desquels règne une grande solidarité, ont un comportement particulier. Nous en avons relevé une centaine, comme *d'un bout_à l'autre*. Sur un plan pratique, il faut les intégrer dans un logiciel de conversion orthographique-phonémique (Catach, 1984).

5. LES LIAISONS ET LES SYNTHESES VOCALES

5.1. Matériel

Voyons à présent ce qui se passe sur un corpus. Nous disposons d'un corpus de 5340 phrases du journal *Le Monde*, chacune étant lue par un nombre de personnes allant de 1 à 10, et transcrite telle qu'elle a été prononcée, par plusieurs experts : BREF 80 (Lamel *et al.*, 1991). Cependant, nous nous sommes limité aux phrases lues par au moins 3 personnes : au

nombre de 143 (soit 2959 mots), elles présentent 315 contextes consonne latente + espace ou trait d'union + voyelle ou *h* - que nous désignerons désormais par CL, tandis que BREF 90 désignera ce nouveau corpus.

Seules 19 phrases, lues par 3 personnes, ne manifestaient pas de CL. Si l'on pondère par le nombre de locuteurs, on a au total 469 liaisons réalisées, sur 1232 CL : les mots déterminatifs (articles, adjectifs démonstratifs et possessifs), quelques autres adjectifs antéposés, certains adverbes, pronoms et prépositions monosyllabiques produisent des liaisons toujours réalisées.

5.2. Les règles des synthèses vocales

Comparons maintenant le traitement des liaisons en français par 5 synthèses vocales à partir du texte : GRAPHON, du LIMSI-CNRS (Prouts, 1980) ; SYNTHE III (Morel), commercialisée par ELECTREL ; VoxBox, du KTH (Carlson *et al.*, 1982), commercialisée par Infovox ; TELEVOX, du CNET (Divay & Guyomard, 1977), commercialisée par ELAN Informatique ; et Apollo II, d'IBM (Dolphin).

GRAPHON est le seul convertisseur graphème-phonème à faire des liaisons en /R/, le seul à faire la liaisons avec *divin*, *plein*, *vain*, mais est aussi le seul à ne pas la matérialiser dans *quant*. Il connaît 71 règles sur les liaisons, auxquelles il faut ajouter quelques unes sur les nombres et sur le *h* aspiré. La section réservée aux liaisons est essentiellement la liste des mots à liaison potentielle, avec leur transcription phonémique, ce qui résout automatiquement les problèmes de dénasalisation, d'aperture et de (dé)voisement. Certaines empêchent la liaison amont - de droite à gauche - (ex. *envers*, *hors*, *hormis*). De plus, sont activées :

(i) une liaison en /z/ entre 2 mots terminés par 's' ou 'x' - sauf après *vers* (ex. *vers*|*eux*) et après l'un des 3 mots cités ci-dessus -,

(ii) une liaison en /t/ (resp. en /z/) après un mot finissant en 'd' ou 't' (resp. en 's', 'x' ou 'z'), séparé du suivant par un trait d'union,

si le 2e mot commence par une voyelle ou un *h* muet.

SYNTHE III ignore (i), mais comprend, parmi ses règles, 18 chaînes de caractère terminées par *-s* ou *-x* qui ne figuraient pas dans les règles de GRAPHON (ex. *jeunes*, mots en *-eux* et en *-aux*). A noter la même incohérence que dans TELEVOX : si les formes masculines de *grand* donnent des liaisons, il n'en va pas ainsi du pluriel *grandes*. SYNTHE III manipule 64 règles sur les liaisons - dont 2 pour *bon(s) à*. Contrairement aux 4 autres synthèses, celle-ci ne réalise pas de liaison pour *petit* (non suivi d'un trait d'union) et *quels*. Par défaut, la liaison est faite après - voyelle + *s/t* + trait d'union, après *certain* (mais sans dénasalisation), après *vingts* (mais en /t/). Seul *yeux*, parmi les mots commençant par *y-*, permet la liaison amont.

Il est moins évident d'attribuer à VoxBox un nombre strict de règles pour les liaisons, car divers points de vue sont combinés (*cf.* § 5.3.) : on peut néanmoins avancer le chiffre minimum de 109 - dont 29 formes de l'auxiliaire *être* (ex. *étai(en)t*, *serais*). Contrairement aux 4 autres synthèses, VoxBox ne fait la liaison après aucune forme du verbe auxiliaire *avoir*, ni après *rien* - à moins qu'un trait d'union ne suive. Elle n'ouvre et ne dénasalise jamais la voyelle précédant la consonne latente, ne dévoise pas le *-d* de *grand* et ne voise pas *-s* suivi d'un trait d'union.

TELEVOX admet moins de règles (67) : ainsi, contrairement aux 4 autres synthèses, celle-ci n'associe de liaison ni à *bon*, ni aux adverbes monosyllabiques *pas*, *plus*, ni aux adjectifs possessifs des personnes du pluriel (*nos*, *vos*, *leurs*). Elle ne force pas les liaisons entre 2 mots séparés par un trait d'union, d'où des erreurs telles que **sait-on*. Les mots en *y-* interdisent la liaison amont, d'où aussi des erreurs comme **ils*|*y vont*.

Avec Apollo II, les mots en *-z* (non en *-ez*), *-x*, *-s*, *-t*, *-d*, *-p*, *-n* (non en *-in*) engendrent par défaut des liaisons aval - exception : *et*. Si *un*, *en* et les formes en *-ien* ne sont pas dénasalisées (ce qui produit entre autres une erreur pour *ancien*, qui, en liaison comme au féminin, est lu */ɑ̃sjɛ̃n/), Apollo II dénasalise

- les autres mots en *-un*, en /-œn/ (ex. *aucun*),
- les autres mots en *-an* et *-en*, en /-an/,
- et ceux en *-on*, en /-ɔn/ (y compris *on*).

En calculant l'intersection entre ces différentes règles, on aboutit à 34 mots - tous monosyllabiques -, pouvant produire une liaison aval universellement partagée par ces 5 synthèses vocales.

5.3. Étude comparative

Par la suite et dans le tableau donné à la fin (table 1), LO (resp. LI) représente les liaisons effectuées (resp. évitées) par la totalité des locuteurs de BREF 90 - LF représentant les autres liaisons (facultatives). Tandis qu'Apollo II réalise plus de 50 % de LI - telle **Mitterrand et* (liaison que SYNTHE III fait aussi) -, TELEVOX est la synthèse qui fait le moins de liaisons - moins encore que VoxBox (la plus proche des locuteurs) et que les sujets eux-mêmes.

Nous avons relevé 161 cas de liaisons pour lesquels existe au moins une divergence entre les synthèses vocales. Plus de la moitié des différences pour les LI et les LO, entre BREF 90 et GRAPHON, viennent de la règle spéciale (i) : ex. *trois mois_après* (à cause de la terminaison en *-s* du dernier mot), *la publication de statistiques_américaines* - liaison que, suivant une autre stratégie, VoxBox ne fait pas (*cf. infra*).

A 2 cas près, les différences entre les résultats de BREF 90 et ceux de VoxBox, pour les LI - toutes en /-Rz/ - et les LO, proviennent d'heuristiques syntaxiques et suprasegmentales de cette synthèse :

- *pas, l'un, tout* précédé d'une pause n'engendrent pas de liaison devant une préposition ;
- *tant* non plus, devant un pronom personnel;
- un mot plein en *-s* ou en *-x* précédé d'un mot déterminatif pluriel engendre une liaison aval si le mot suivant n'a pas une terminaison identifiée comme verbale (ex. *-issions*): d'où, contrairement à ce que donne BREF 90, *des renforts_envoyés, des nouvelles émissions.*

Contrairement à ce qui se passe dans BREF 90 et les synthèses précédentes, avec TELEVOX, SYNTHE III et Apollo II, les pronoms accentués *certains* et *on* inversé engendrent des liaisons aval, et on a des prononciations telles que */ateil/ pour *a-t-il*. Il faut ajouter que VoxBox et SYNTHE III, à la différence des autres synthèses, peuvent engendrer une liaison amont avec un 'w'. Pourtant, la semi-voyelle (ou semi-consonne) /w/ ne permet jamais la liaison lorsqu'elle est transcrite par 'w' (ex. *un whisky*). De plus, *oui* et *ouistiti* se comportent comme s'ils commençaient par une consonne (ex. *un oui franc et massif*).

6. CONCLUSION

Nous avons défini la liaison, comment et quand on la fait - en nous en tenant aux liaisons obligatoires et illi-cites -, avant d'étudier son traitement par 5 synthèses vocales.

Les liaisons sont souvent optionnelles, parfois inattendues, ce qui peut contribuer à l'impression d'un style ou d'un autre : outre l'histoire et l'esthétique, les paramètres idiolectaux, les habitudes langagières rendent compte d'écarts non négligeables. Il faut donc rester circonspect : on ne peut pas développer de certitudes dans ce domaine, même si les grammairiens ne manquent pas, pour proclamer ce qui est juste, légitime, convenable en "bon français", ou au contraire ce qui est une faute, ce qui n'est pas correct, pas académique. Ce fait est étayé par la terminologie fondamentalement normative qui est en usage - et que nous avons été amené à adopter.

7. BIBLIOGRAPHIE

Carlson R., Granström B. & Hunnicutt S. (1982) "A multi-language text-to-speech module", *Proc. IEEE-ICASSP*, vol. 3, Paris.

Catach N. (1984) *La phonétisation automatique du français, les ambiguïtés de la langue écrite*, éditions du CNRS, Paris.

Divay M. & Guyomard M. (1977) *Conception et réalisation sur ordinateur d'un programme de transcription graphémo-phonétique du français*, Thèse de Troisième Cycle, Université de Rennes.

Donohue-Gaudet M.-L. (1963) *Le vocalisme et le consonantisme français*, Librairie Delagrave, Paris.

Eggs E. & Mordellet I. (1990) *Phonétique et phonologie du français, Théorie et pratique*, Niemeyer Verlag, Tübingen.

Encrevé P. (1988) *La liaison avec et sans enchaînement. Phonologie tridimensionnelle et usages du français*, Éditions du Seuil, Paris.

Fouché P. (1969) *Traité de prononciation française*, Éditions Klincksieck, Paris.

Lamel L.F., Gauvain J.-L. & Eskénazi M. (1991) "BREF, a Large Vocabulary Spoken Corpus for French", *Proc. EUROSPEECH*.

Léon P.R. (1976) *Introduction à la phonétique corrective à l'usage des professeurs de français à l'étranger*, Hachette/Larousse, Paris.

Lucci V. (1983) *Étude phonétique du français contemporain à travers la variation situationnelle*, Publications de l'Université des Langues et Lettres de Grenoble.

Pérennou G. & Calmès M. de (1987) *BDLEX, base de données lexicales du français écrit et parlé*, Travaux du laboratoire CERFIA, Toulouse.

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 de Paris Sud.

Wauquier-Gravelines S. (1994) "Segmentation lexicale en français parlé. La liaison enchainée", *XX^es JEP*, Trégastel.

Table 1 : liaisons "obligatoires", "facultatives" et "interdites", omises ou réalisées par 5 synthèses vocales.

%	GRAPHON	SYNTHE III	VoxBox	TELEVOX	Apollo II
LF réalisées	63,5	61,2	52,6	34,3	89,2
LO omises	0,8	2,8	4,4	8,3	2,8
LI réalisées	5,1	10,6	3,0	2,0	53,9

RÔLE DES CHANGEMENTS DE LA DURÉE ET DE L'INTENSITÉ DANS LA SYNTHÈSE DU TCHÈQUE

Marie DOHALSKÁ-ZICHOVÁ, Tomáš DUBĚDA

Institut de Phonétique, Faculté des Lettres, Université Charles, Prague
nám. Jana Palacha 2, 116 38 Prague 1, République Tchèque
Tél.: 21619 250, Fax.: 744458, e-mail: marie.dohalska@ff.cuni.cz

ABSTRACT

The present article deals with our recent research concerning the role of duration (*t*) and intensity (*i*) in the diphone TTS synthesis of Czech. The regularities that became evident: 1) In the last stress unit of a sentence, *t* increases and *i* decreases. 2) Not infrequently, a slight shortening of the stress syllable suggest a better prosodic effect. These rules work despite the phonological distinction between short and long vowels in Czech. We have already applied the preliminary values on a sample corpus, and we verified them in following perception tests. The introduction of this knowledge into automatic text-to-speech rules (which neglect *i* fully and *t* partly, at the present stage) could lead to considerable quality improvement of the synthesis.

1. INTRODUCTION

L'état actuel de la synthèse par diphones à partir du texte de la langue tchèque est caractérisé par des règles suprasegmentales restreintes:

- aux changements de la **fréquence fondamentale,**

- à une simple modification de la **durée** des diphones en fonction de la longueur du mot. (Plus l'unité accentuelle est longue, plus la durée des diphones est réduite. La réduction est répartie sur tous les diphones de l'unité accentuelle).

L'**intensité** est totalement négligée.

Nos études récentes de la parole spontanée de différents débits ont révélé certaines régularités des changements de la **durée** *(t)* et de l'**intensité** *(i)* dans les schémas prosodiques de base, surtout à la fin de la phrase. L'incorporation de ces deux paramètres devrait mener à une amélioration sensible de la synthèse du tchèque.

Dans nos essais de modélisation manuelle, il s'est avéré que les changements de *t* et *i* sont susceptibles d'introduire une impression d'une **courbe mélodique naturelle** de la phrase tchèque.

2. LA "PHRASE MODÈLE"

Tout d'abord, nous avons créé 12 **phrases modèles** les plus satisfaisantes du point de vue de la perception, à l'aide de la correction manuelle de la fréquence fondamentale, de la durée et de l'intensité des segments. Les corrections ont été basées sur nos analyses précédentes de la parole naturelle (corpus d'une durée de 160 minutes). Les valeurs des trois paramètres (*f, t* et *i*) ont été précisées pour chaque son.

Les phrases étaient d'une longueur de 12 - 16 syllabes. Puisqu'il s'agissait d'une première approche, nous avons consciemment choisi un corpus très limité.

Voici 3 phrases de notre corpus:*

*1. Dnes končí // počáteční / fáze / **výzkumu.***

*2. Dnes končí // počáteční / fáze / **letu.***

*3. Dnes končí // počáteční / fáze / **žní.***

*(1. Aujourd'hui, on achève la phase initiale **de la recherche.***

*2. Aujourd'hui, on achève la phase initiale **du vol.***

*3. Aujourd'hui, on achève la phase initiale **de la moisson.**)*

*** Segmentation prosodique en tchèque:**

/ ... frontière de l'unité accentuelle (très souvent identique à un mot, en tchèque)

// ... frontière majeure (sémantique)

Les graphiques présentés en annexe nous permettent de suivre la répartition **temporelle** de trois versions de la phrase *Dnes*

končí počáteční fáze letu (à gauche) et les valeurs de l'**intensité** correspondantes (à droite).

Devant les frontières majeures à l'intérieur de la phrase, on remarque une légère diminution de l'intensité et un ralentissement (augmentation de *t*) faible; pour l'unité accentuelle finale (notamment pour sa syllabe finale), ces deux tendances sont encore plus renforcées. Les deux paramètres *t* et *i* se comportent donc de façon opposée.

En plus, nous avons constaté que ces tendances sont valables aussi bien pour les voyelles brèves que pour les voyelles longues, malgré leur opposition dans le système phonologique de la langue tchèque.

Dans les tables suivantes, nous présentons les valeurs moyennes de *t* et *i* tirées des phrases modèles, en fonction de la position du son dans la phrase. (Il est évident que les valeurs sont basées, jusqu'ici, sur un corpus restreint.)

Les valeurs de *t* et *i* dans les tables présentent les pourcentages par rapport aux valeurs de base. Il s'agit donc de données relatives.

Table 1: Durée et intensité dans l'unité accentuelle **non-finale** (valeurs moyennes)

Syllabe initiale		Syllabe finale	
t	*i*	*t*	*i*
96%	111%	104%	97%

Dans l'unité accentuelle non-finale (table 1), on remarque une hausse de *t* (4% au-dessous de la moyenne dans la syllabe initiale, 4% au-dessus de la moyenne dans la syllabe finale), un *i* relativement haut dans la syllabe initiale (celle-ci porte toujours l'accent en tchèque), et une baisse légère de l'intensité dans la syllabe finale.

Table 2: Durée et intensité dans l'unité accentuelle **finale trisyllabique** (valeurs moyennes)

	1ère syllabe	2e syllabe	3e syllabe
t	100%	100%	150%
i	113%	85%	54%

Dans l'unité accentuelle finale trisyllabique (table 2), l'allongement ne frappe que la dernière syllabe. La chute de *i* est graduelle.

Table 3: Durée et intensité dans l'unité accentuelle **finale bisyllabique** (valeurs moyennes)

	1ère syllabe	2e syllabe
t	106%	138%
i	106%	61%

Table 4: Durée et intensité dans le groupe rythmique **final monosyllabique** (valeurs moyennes)

	1ère syllabe
t	126%
i	70%
chute de *i*	50%

Dans la table 4, l'allongement de *t* et la baisse graduelle de *i* se fait sur une seule syllabe, d'où la chute brusque de *i*.

Dans toutes les tables, nous prenons en considération uniquement les syllabes contenant les voyelles brèves. En ce qui concerne les voyelles longues, nous mentionnons, à titre d'exemple, quelques données:

Table 5: Durée des **voyelles longues** (dans toutes les positions - valeurs moyennes)

Mot trisyllabique - 1ère syllabe	300%
Mot trisyllabique - 3e syllabe	369%
Mot monosyllabique	370%

3. TEST D'ÉCOUTE

Pour vérifier nos hypothèses, nous avons préparé plusieurs tests préliminaires, dans lesquels nous avons comparé différentes modifications des phrases modèles, pour examiner l'importance de chacun des paramètres (*t*, *i* et *f*).

Nous citons ici, à titre d'exemple, les résultats d'un des tests réalisés:

D'abord, nous avons élaboré trois modifications pour chaque phrase:

A - phrase modèle **sans changement de durée** (la durée de chaque son de la phrase modèle a été égalisée à 100%).

B - phrase modèle **sans changement de l'intensité** (l'intensité de chaque son de la phrase modèle a été égalisée à 100%).

C - phrase modèle avec la courbe mé-

lodique **extraite de la synthèse automatique** actuelle (valeurs de *t* et *i* „manuels" pour chaque son, valeurs de *f* générées par le système de la synthèse automatique).

Nous avons présenté ces trois variantes, dans un ordre aléatoire, à un groupe de 37 personnes (étudiants de divers spécialisations), avec la seule instruction de choisir la phrase la plus satisfaisante du point de vue de la perception.

3.1. Les résultats:

Table 6: Résultats du test d'écoute

Variante	Préférence (%)
A - sans changements de la durée	41
B - sans changements de l'intensité	43
C - avec la fréquence automatique	16

3.2. Interprétation des résultats du test

1) Les variantes **A** (sans changements de *t*) et **B** (sans changement *i*) ont un pourcentage très similaire et assez élevé, ce qui affirme l'importance des deux paramètres *t* et *i* pour la qualité prosodique de la phrase synthétique tchèque.

2) La variante **C** (avec la fréquence automatique) a atteint, malgré son caractère moins naturel, un pourcentage non négligeable de 16% qui nous semblait relativement haut. Cela peut être dû à la présence de la durée et de l'intensité, qui ont, en quelque sorte, compensé la fréquence moins parfaite.

4. CONCLUSION

Dans la parole naturelle, les trois paramètres (*t, i* et *f*) sont toujours inséparablement liés l'un à l'autre. Par souci de simplification, nous les avons traités séparément. C'est grâce à cette méthode que nous avons pu entreprendre les premiers pas dans le déchiffrage des problèmes prosodiques de la synthèse par diphones actuelle.

Le principe de la baisse d'intensité et du ralentissement graduels sur différents niveaux prosodiques (unité accentuelle, unité mélodique, phrase) est conditionné, en partie, par la physiologie de la parole. Il est connu et appliqué depuis longtemps dans un grand nombre de langues. En tchèque, il est accepté avec modération, pour les raisons suivantes:

- il est en contradiction partielle avec l'opposition phonologique des voyelles brèves et des voyelles longues,
- un allongement excessif en fin de phrase peut donner l'impression d'une prononciation familière.

Il résulte de nos sondages

- qu'il serait souhaitable d'enrichir le système de synthèse automatique par les paramètres *t* et *i*,
- que la durée des voyelles obéit aussi bien aux lois phonologiques (opposition brève - longue) qu'aux lois prosodiques.

Nous nous concentrerons maintenant sur les problèmes suivants:

- automatisation des trois paramètres,
- vérification des règles formulées sur un corpus plus riche et plus varié,
- incorporation des résultats dans le système de la synthèse automatique.

5. BIBLIOGRAPHIE

Dohalská-Zichová M. (1991) La vitesse d'articula-tion et les unités sonores dans la chaîne parlée, *Actes du XIIe ICPhS*, 390 - 393

Dohalská M. - Ptáček M. (1994) Quelques remarques sur la perception du tchèque synthètique, *AUC Phonetica Pragensia VIII,* 79-126

Hess W. - Kraft V. - Portele T. (1994) Zum Problem der Evaluierung von Sprachsynthesesystemen - dargestellt am Beispiel der Synthesekomponenten in Verbmobil, *XX. Deutsche Jahrestagung für Akustik - DAGA 94,* Vol A, 103-116

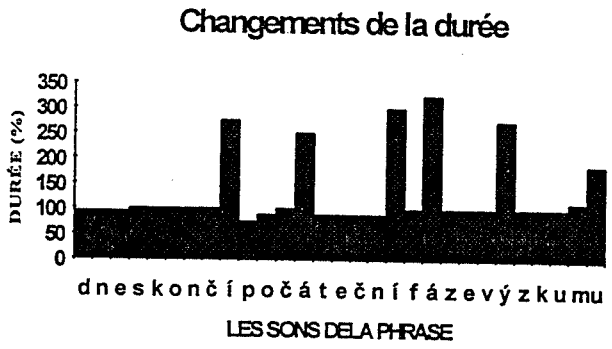
Ptáček M. - Vích R. - Víchová E. (1992) Czech Text-to-Speech Synthesis by Concatenation of Parametric Units, *URSI, Signals Systems and Electronics,* 230-232

Janota P. - Dohalská M. - Palková Z. - Ptáček M. (1994) Current Situation in the Research of Automatic Generation of the Prosodic Features with the Diphone Synthesis of Spoken Czech, *AUC Phonetica Prague VIII, 33 - 58*

Rossi M. - Di Cristo A. (1982) En quête des indices de segmentation prosodiques de l'énoncé, *Actes du Séminaire „Prosodie et Reconnaissance automatique de la parole", Aix-en-Provence, 7 - 8 octobre 1982, 141 - 164*

ANNEXE: Graphiques

Fig. 1 et 2: Phrase *Dnes končí // počáteční / fáze / výzkumu. (Aujourdhui, on achève la phase initiale de la recherche.)*



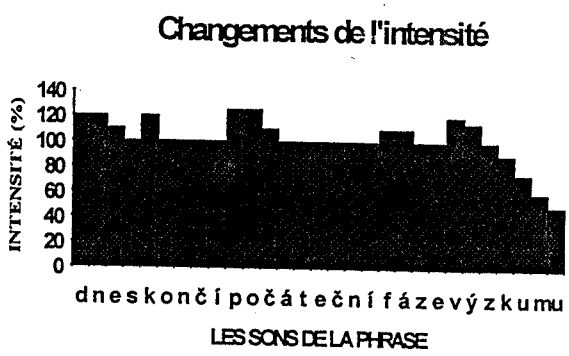


Fig. 3 et 4: Phrase *Dnes / končí // počáteční / fáze / letu. (Aujourd'hui, on achève la phase initiale du vol.)*

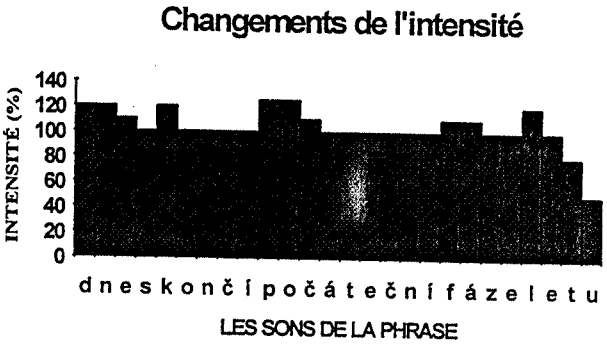
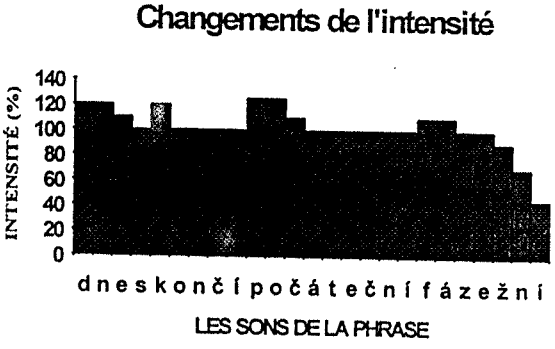


Fig. 5 et 6: Phrase *Dnes končí // počáteční / fáze / žní. (Aujourd'hui, on achève la phrase initiale de la moisson.)*



SYNTHESE AUDIOVISUELLE DE LA PAROLE A PARTIR DU TEXTE

Bertrand LE GOFF, Christian BENOIT

Institut de la Communication Parlée, BP 25X, 38040 Grenoble Cedex 9, France

Tél. : 76 57 76 45 – Fax : 76 82 43 35 – e-mail: legoff@icp.grenet.fr, benoit@icp.grenet.fr

ABSTRACT

There is valuable and effective information afforded by a view of the speaker's face in speech perception and recognition by humans. Visible speech is particularly effective when the auditory speech is degraded, because of noise, bandwidth filtering, or hearing-impairment. There is evidence that synthetic faces increase the intelligibility of synthetic speech when the facial gestures and speech sounds are coherent. To reach this goal, the articulatory parameters of the facial animation have to signal the same message as the auditory speech. If a completely ambiguous message or even a contradictory message is given by the visible speech, then intelligibility would necessarily be decreased. Most of the existing parametric models of the human face have been developed in the perspective of optimising the visual rendering of facial expressions. As regards visual speech synthesis, most of the existing systems are based on a limited set of facial images occurring in the natural production of speech that are displayed one after the other, depending on the phoneme -- or rather "viseme" -- simulataneously uttered by the acoustic synthesizer. Actually, the coarticulation effects and the transition smoothing are much more naturally simulated by means of parametric models specially controlled for visual speech animation. In that perspective, a high-resolution model of the lips has been developed at the ICP. It is controlled by only five parameters. It has been implemented onto the face model originally developed by Parke and specially modified so that it is now controlled by eight articulatory parameters; the tongue and the chin motions being also rendered. Following the strategy adopted by Cohen and Massaro (1993), we have assigned target values and dominance functions to each of the eight parameters caracteristic of a French viseme. This technique allows the evolution of the parameters to be smoothed over time and gives a fair account of the coarticulation phenomenon. Our visual synthesizer has finally been synchronized with the ICP diphone-based Speech Synthesizer.

1. INTRODUCTION

Les informations visuelles apportées par le visage de l'interlocuteur contribuent de façon importante à l'amélioration de l'intelligibilité de la parole. La parole visuelle est particulièrement utile lorsque le signal auditif est dégradé (Sumby et Pollack, 1954; Erber, 1975; Massaro, 1987; Summerfield et al., 1989; Benoît et al., 1994), mais son influence ne se limite pas seulement aux situations d'audio bruité. La vidéo d'un locuteur articulant "aga" et doublée auditivement par "aba" sera perçue audiovisuellement comme "ada" (McGurk et MacDonald, 1976).

Dans le but d'améliorer l'interface homme-machine, les ordinateurs ont été dotés de voix de synthèse sans pour autant avoir un visage. La voix de synthèse n'ayant pas encore les caractéristiques de la voix humaine, elle a alors besoin d'un apport supplémentaire d'intelligibilité. Des études ont montré qu'un visage naturel pouvait reconstituer les deux-tiers de l'intelligibilité manquante dans une situation de communication où l'acoustique est dégradée (Benoît et al., 1994). Un visage synthétique en restaure la moitié (Le Goff et al., 1995). Une synthèse audiovisuelle doit donc être plus intelligible qu'une simple synthèse acoustique, à condition toutefois que le message linguistique synthétisé soit cohérent entre les deux modalités. Une telle exigence ne peut être obtenue que grâce à un modèle paramétrique de visage correctement contrôlé de façon à restituer la coarticulation naturelle de la parole. En effet, les systèmes basés sur la concaténation d'images-cibles préstockées n'ont pas l'adaptabilité nécessaire pour atteindre cet objectif. A notre connaissance, il n'existe à ce jour que trois synthétiseurs audiovisuels de la parole à partir du texte basés sur des modèles tridimensionnels de visages paramétriques : ceux d'UCSC, de l'ICP et du KTH (Benoît et al., 1995).

2. LE SYNTHETISEUR DE VISAGE PARLANT

Notre synthétiseur de visage parlant est basé sur un modèle tridimensionnel de visage à huit paramètres. Une vingtaine de "visèmes" (*visual phoneme*, Fisher, 1968) étant nécessaires pour décrire le jeu symbolique des lèvres en français (Benoît et al., 1992), un octuplet de valeurs-cible a été déterminé à partir de mesures labiométriques pour chacun de ces visèmes. Enfin, les trajectoires des paramètres "entre" ces cibles (pas forcément atteintes) sont calculées par des règles mathématiques à partir de fonctions de dominance définies par trois valeurs (intensité de l'attraction, amplitudes temporelles gauche et droite) pour un paramètre donné et un visème donné. Ces valeurs ont été calculées à partir d'observations sur un locuteur français.

Notre modèle de visage est donc animé à partir d'une chaîne de visèmes définis par leur durée. Cette chaîne de visèmes est, en particulier, facile à obtenir, par un jeu simple de réécriture, à partir d'une chaîne phonétique marquée en durée fournie par un synthétiseur acoustique.

2.1. Le modèle de visage

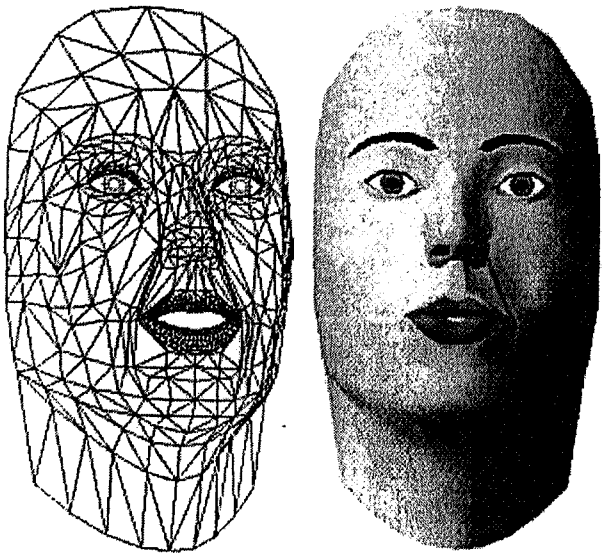


Figure 1: Modèle de visage de l'ICP. A gauche, structure fil-de-fer sans les dents ni la langue prononçant un /i/ stable. A droite, rendu par méthode Gouraud prononçant un /y/ stable.

Le modèle de visage que nous avons implanté est le résultat de la greffe du modèle tridimensionnel de lèvres élaboré à l'ICP (Guiard-Marigny et al., 1994) sur le modèle de visage originalement développé par Parke (1974) et modifié par Cohen et Massaro (1990), puis par Le Goff (1993).

Nous pilotons ce modèle à l'aide de huit paramètres, dont cinq pour les lèvres (écartement horizontal interne, séparation verticale interne, protrusion du point de contact des lèvres C, protrusion des lèvres inférieure et supérieure), un pour la position du menton (i.e. de la mâchoire), et deux pour la langue (angle de la langue avec la mâchoire, avancement de la langue).

2.2. Valeurs-cible des paramètres

Les valeurs-cible sont les valeurs qui étaient atteintes par les paramètres sans coarticulation. Ces valeurs-cible ont été mesurées sur des voyelles tenues et sur une moyenne des observations de chaque consonne dans les trois contextes i, a, y.

Nous utilisons dans cette première version dix-neuf catégories de "visèmes" : [a], [i], [y, u, ø, o, õ], [e, ɛ, ɛ̃], [ɔ], [œ, œ̃], [ã], [p, b, m], [t, d, n], [f, v], [s, z], [ʃ, ʒ], [l], [ʁ], [w], [j], forme préphonatoire (lèvres entrouvertes), forme de repos (lèvres fermées).

Ces valeurs-cible ont été mesurées à partir de mesures effectuées sur un locuteur filmé de face et de profil.

2.3. Fonction de dominance

La fonction de dominance d'un visème représente l'influence de ce visème sur ses voisins. Elle se définit à l'aide de trois coefficients (α, θ_1, θ_2) pour chacun des huit paramètres de commande : nous utilisons donc vingt-quatre coefficients. Le coefficient α_{ij} est l'intensité absolue d'attraction du visème i pour le paramètre j. Le coefficient $\theta_{1_{ij}}$ est le reflet de la croissance de la fonction : c'est la coarticulation anticipatrice. Le coefficient $\theta_{2_{ij}}$ est le reflet de la décroissance de la fonction : c'est la coarticulation rétentive. Les vingt-quatre coefficients ont été finalement obtenus par ajustements successifs de façon à approcher au plus près l'ensemble des mesures effectuées sur un locuteur.

La fonction de dominance que nous avons utilisée est l'une de celles proposées par Cohen et Massaro (1993), de la forme

$f_{ij}(t) = \alpha_{ij}.e^{-\theta_{ij}|t_0-t|}$, où t_0 correspond au centre acoustique de la réalisation phonétique correspondante.

L'évolution temporelle du paramètre j prend alors une allure sigmoïdale entre deux cibles, dont la forme dépend des coefficients caractéristiques de chaque cible.

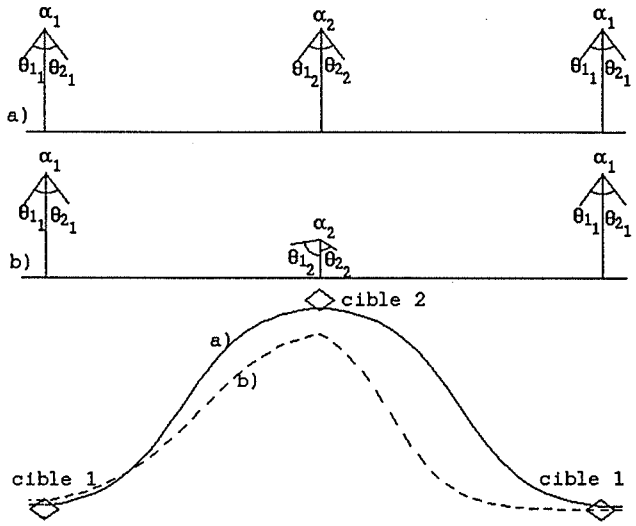


Figure 2: En haut, fonction de dominance et coefficients; en bas, trajectoire théorique d'un paramètre entre trois cibles (représentées par des losanges) Cas a) de dominances identiques. Cas b) de dominances différentes.

Sur la figure 2, nous avons représenté la trajectoire théorique d'un paramètre entre 3 valeurs-cible. Le premier cas, noté a), en trait plein, est celui de cibles de dominance identique. On voit que la cible 2 est alors atteinte. Le deuxième cas, noté b), en trait pointillé, montre la trajectoire d'un paramètre qui n'atteint pas la cible 2 parce que α_2 est trop petit. On observe donc un *"undershoot"*. Ce second exemple présente également une forte coarticulation anticipatrice (θ_{12} grand) et une faible coarticulation rétentive (θ_{22} petit) qui dissymétrise clairement la trajectoire. Ce calcul mathématique est complété par une correction de trajectoire sur le paramètre de séparation verticale afin que les fermetures des occlusives bilabiales et des fricatives labio-dentales soient effectives.

2.4. Fonctionnement du système

Les synthèses acoustique et visuelle s'effectuent en parallèle grâce au logiciel PVM (*Parallel Virtual Machine*).

La synthèse acoustique, créée par Bailly et Guerti (1991), est réalisée par concaténation d'unités stockées (polysons). L'ensemble des règles permettant la synthèse acoustique a été réalisé sous le formalisme COMPOST (Bailly et Tran, 1989). Ce COMpilateur Phonétique sur Ordinateur pour la Synthèse de Texte manipule des règles et des objets via la manipulation de listes d'atomes, d'arbres et de sous-arbres.

COMPOST analyse l'entrée (texte orthographique français) et fournit, entre autres, une chaîne de phonèmes avec leur durée, seules informations utilisées pour le moment dans notre synthétiseur de visage parlant.

Parallèlement, la synthèse visuelle effectue une conversion phonème/visème. Les trajectoires sont calculées à partir des cibles correspondant aux visèmes de la phrase. Les paramètres sont finalement stockés dans un fichier.

En sortie de synthèse, les deux fichiers visuels et acoustiques sont utilisés pour l'animation du visage synthétique avec la même technique que celle utilisée pour l'analyse-synthèse de visages réels.

3. CONCLUSION

Nous avons présenté un premier prototype de synthétiseur de la parole audiovisuelle à partir du texte. Il fonctionne en français à partir d'une entrée orthographique quelconque. Ce système a de multiples applications, notamment dans les interfaces homme-machine, l'apprentissage du français langue étrangère, ou l'apprentissage de la lecture labiale. Cependant, ce n'est qu'une première étape vers un système plus complet. Il reste à affiner les fonctions de dominance à l'aide d'un corpus plus étendu. De même, les expressions faciales, telles que clignement et mouvements d'yeux, hochement de tête, mouvement de sourcils seront implémentées et commandées par la prosodie et la syntaxe. Enfin, des tests sont en cours de réalisation afin d'évaluer l'intelligibilité de cette synthèse par rapport à un locuteur naturel.

Remerciements : Cette recherche a bénéficié du soutien du Ministère de l'Enseignement Supérieur (Bourse doctorale du MESR) et du projet européen ESPRIT-BRA N° 8579 "MIAMI".

4. BIBLIOGRAPHIE

Bailly G. et Tran A. (1989) Compost: A rule-compiler for speech synthesis, *Proceedings of Eurospeech Conference*, Paris, France, 136-139.

Bailly G. et Guerti M.(1991) Synthesis-by-rule for French, *Proceedings of the XIIth International Congress of Phonetic Sciences*, Aix-en-Provence, France, n°2, 506-511.

Benoît C., Lallouache M.T., Mohamadi T. et Abry C. (1992) A set of French visemes for visual speech synthesis, *Talking Machines: Theories, Models and Designs*, G. Bailly and C. Benoît, Eds, Elsevier Science Publishers B.V., North-Holland, Amsterdam, 485-504.

Benoît C., Mohamadi T. et Kandel S. (1994) Audio-Visual Intelligibility of French speech in noise, *Journal of Speech & Hearing Research*, n°37, 1195-1203.

Benoît C., Beskow J., Cohen M., Granstrom B., Le Goff B. et Massaro D.W. (1995) Text-to-Audio-Visual Speech Synthesis over the world, *Proceedings of the Speech Maps workshop*, C. Abry et P. Badin Eds., Grenoble, 5-6 décembre 1995.

Cohen M.M. et Massaro D.W. (1990) Synthesis of visible speech, *Behaviour Research Methods, Instruments \& Computers*, 22(2), 260-263.

Cohen M.M. et Massaro D.W. (1993) Modeling coarticulation in synthetic visual speech, *Proceedings of Computer Animation93*, Magnenat-Thalmann & Thalmann Eds, Geneve, Suisse.

Erber N.P. (1975) Auditory-visual perception of speech, *Journal of Speech & Hearing Disorders*, n°40, 481-492.

Fisher C.G. (1968) Confusions among visually perceived consonants, *Journal of Speech & Hearing Research*, n°15, 474-482.

Guiard-Marigny T., Adjoudani A. et Benoît C. (1994) A 3D model of the lips, *Proceedings of the 2nd ETRW on Speech Synthesis*, New Platz, USA.

Le Goff B. (1993) Commandes paramétriques d'un modèle de visage 3D pour animation en temps réel, *Mémoire de D.E.A. Signal Image Parole*, Institut National Polytechnique, Grenoble, France, 106 pp.

Le Goff B., Guiard-Marigny T. et Benoît C.(1995) Read my lips ... and my jaw! How intelligible are the components of a speaker's face? *Procedings of Eurospeech'95*, Madrid, Spain.

McGurk H. et MacDonald J. (1976) Hearing Lips and Seeing Voices, *Nature*, 264, 746--748.

Massaro D.W. (1987) Speech perception by ear and eye: a paradigm for psychological inquiry., *LEA*, Hillsdale, NJ.

Parke F.I. (1974) A parametric model for human faces, *PhD Dissertation*, University of Utah, Department of Computer Sciences.

Sumby W.H. et Pollack I. (1954) Visual contribution to speech intelligibility in noise, *Journal of the Acoustical Society of America*, n°26, 212-215.

Summerfield Q., MacLeod A., McGrath M. et Brooke M. (1989) Lips, teeth, and the benefits of lipreading, in *Handbook of Research on Face Processing*, A.W. Young & H.D. Ellis Editors, Elsevier Science Publishers, 223-233.

Utilisation de techniques d'apprentissage automatique pour les traitements linguistiques et prosodiques en synthèse de la parole : quelques résultats en Anglais, Allemand et Français.

Boëffard, O, Bigorgne, D., Cherbonnel, B., Emerard, F., Roussarie, L., Bagshaw, P., Conkie, A., Ennilo, M., Traber, C.

France Telecom - CNET, LAA/TSS/RCP - 2 av Pierre Marzin, 22307 Lannion Cedex

Abstract

This paper presents the use of automatic learning procedures in both linguistic and prosodic domains for the CNET multilingual text-to-speech system. Using automatically segmented and labelled databases, this approach is applied to the tagging and the parsing of English texts, to the modelling of duration for French, and to the generation of F0 contours for German.

1. Introduction

Des efforts importants ont été entrepris depuis ces cinq dernières années pour développer un environnement de synthèse multilingue de manière à ce que les méthodologies de développement des nouvelles langues soient uniformisées tant sur les aspects acoustiques que linguistiques et prosodiques. Par ailleurs, la mise en oeuvre de nouveaux services vocaux utilisant la synthèse de la parole à partir du texte implique une offre de plus en plus diversifiée tant sur les voix de synthèse dans une langue que sur les langues elles-mêmes.

L'objectif de cet article vise essentiellement à faire le point sur l'utilisation actuelle au CNET de techniques d'apprentissage automatique pour les traitements linguistique et prosodique en synthèse de la parole. Trois raisons principales ont guidé ce choix : l'accessibilité à des bases de données de plus en plus volumineuses et de nature variée, les progrès de la reconnaissance de la parole en étiquetage du signal acoustique, enfin l'utilisation efficace et de plus en plus importante des techniques d'apprentissage automatique dans le domaine du traitement de la parole [Hirschberg & Prieto, 1994; Ostendorf & Veilleux, 1994]. Dans un premier paragraphe, les solutions qui ont été retenues pour la réalisation des systèmes de synthèse de la parole du CNET actuellement commercialisés sont présentées. Les deux paragraphes suivants précisent l'application de techniques d'apprentissage automatique pour le développement des traitements linguistiques de l'Anglais, la construction de modèles de prédiction de la durée segmentale en Français et du fondamental en Allemand.

2. Le système de synthèse multilingue du CNET

La plupart des systèmes actuels de synthèse de la parole à partir du texte s'articule autour de trois étages principaux : un étage linguistique qui analyse le message fourni en entrée du système, un étage prosodique qui détermine les patrons mélodiques et rythmiques adéquats, un étage acoustique qui génère le signal de parole synthétique. En ce qui concerne le système de synthèse multilingue du CNET, l'étage acoustique, réalisé par un synthétiseur PSOLA [Hamon & al., 1989] est identique, quelle que soit la langue synthétisée; par contre, les deux étages linguistique et prosodique diffèrent pour chacune des langues mais suivent la même méthodologie de développement. Jusqu'à récemment, les systèmes de synthèse CNET disponibles disposaient tous d'étages linguistique et prosodique à base de règles élaborées par un expert.

2.1 Les traitements linguistiques

L'étiquetage en catégories grammaticales, préalable indispensable à l'opération de découpage syntaxico-prosodique, repose sur la consultation d'un lexique de mots et d'expressions idiomatiques (de l'ordre de 15000 items en Français) et de tables de racines, préfixes et suffixes. La consultation des lexiques et tables associée à un ensemble de règles linguistiques permet d'attribuer à chaque mot l'une des catégories grammaticales prévues pour chaque langue. A l'usage, cette procédure se révèle relativement efficace mais peu satisfaisante dans la mesure où l'expertise nécessaire à l'étude d'une nouvelle langue est très coûteuse.

Comme pour l'étiquetage grammatical, le découpage syntaxico-prosodique, ou parsing, est fondé sur un ensemble de règles élaborées par un expert (en Français, 150 règles de découpage ont été définies pour localiser et hiérarchiser les frontières syntaxico prosodiques); une règle est appliquée quand les catégories grammaticales qui la composent coïncident avec la suite de catégories examinée. Là encore, le système ne peut évoluer que par l'introduction de règles supplémentaires déterminées par un expert.

2.2 Les traitements prosodiques

Jusqu'à ce jour, des bases de données prosodiques de taille relativement modeste (de l'ordre de 6500 segments phonétiques pour le Français) étaient constituées manuellement. Leur exploitation statistique systématique a permis de proposer un certain nombre de contours mélodiques-type pour chaque mot d'une longueur donnée et d'une "localisation syntaxique" donnée [Schnabel, 1988, Emerard & al., 1992]. De la même façon, ce type d'analyse fournit un ensemble de règles qui déterminent la durée segmentale en fonction de paramètres tels que le marqueur syntaxico-prosodique qui gouverne le mot, la position de la syllabe dans le mot auquel appartient le phonème, sa position dans la syllabe, son environnement phonétique,... [Bartkova & Sorin, 1987].

3. Utilisation des techniques d'apprentissage pour le développement des modules linguistiques

La stratégie suivie au CNET pour le développement des nouvelles langues et l'amélioration de traitements déjà existants, qu'ils soient de nature linguistique ou prosodique, consiste à développer des modèles de prédiction fiables et précis dont les paramètres sont estimés par apprentissage automatique à partir d'un ensemble suffisamment représentatif de données. Cette approche est illustrée dans ce paragraphe par la présentation du développement de certains traitements linguistiques de l'Anglais : l'étiquetage grammatical et le découpage syntaxico-prosodique.

3.1 Etiquetage grammatical

L'étiquetage grammatical consiste à affecter à chaque mot du message à synthétiser une ou plusieurs étiquettes de nature grammaticale. L'étiquette grammaticale d'un mot est tout d'abord recherchée dans un lexique de 110000 entrées dont chaque item a été correctement étiqueté par un expert. Les mots qui ne sont pas trouvés dans le lexique de référence sont dotés d'une étiquette par défaut. Un ensemble de "règles lexicales" de nature essentiellement morphologique est appliqué sur ces mots de façon à substituer à l'étiquette par défaut une catégorie statistiquement plus vraisemblable. L'apprentissage de ces règles [Brill, 1993] a été mené sur le lexique de 110000 entrées. Ensuite, des règles contextuelles sont activées pour modifier l'étiquette d'un mot compte tenu de son environnement grammatical. L'apprentissage de ces règles a été conduit sur le corpus du Wall Street Journal [Linguistic Data Consortium (LDC)] . Au total, cette méthode permet d'obtenir un taux d'étiquetage correct de 86% avec la première étiquette et de 94% avec les deux premières étiquettes (il y a en moyenne 1,8 étiquette par mot). L'étiquettage par défaut conduit à 48% d'étiquettes correctes.

3.2 Découpage syntaxico-prosodique

Le découpage syntaxico-prosodique, indispensable aux traitements prosodiques, consiste à découper le message d'entrée, dont les mots sont étiquetés grammaticalement, en groupes syntaxico-prosodiques hiérarchisés. Ce découpage se fait par l'application de règles. Pour l'Anglais, elles ont été acquises par apprentissage automatique sur des phrases du corpus Penn Treebank II [LDC] correctement parenthésées par un expert : pour chaque phrase du corpus d'apprentissage, un parenthésage initial est effectué; les règles sont apprises de façon à transformer ce parenthésage initial en un découpage plus proche de celui effectué par l'expert. Un ensemble de 200 règles a été généré de cette manière; elles offrent un parenthésage correct pour 81% des phrases du corpus de test (la moitié de Penn Treebank II).

4. Utilisation des techniques d'apprentissage pour le développement des modules prosodiques

Offrir une voix de synthèse demeure, aujourd'hui encore, une tâche coûteuse en temps et en expertise à la fois pour constituer un répertoire d'unités et pour élaborer un modèle prosodique adapté à chaque répertoire.

Afin d'accélérer le temps de mise au point et de réduire l'expertise nécessaire, des outils d'automatisation des tâches ont été développés. Ils concernent d'une part la segmentation des répertoires d'unités acoustiques [Boëffard, 1993] et d'autre part, l'extraction automatique de mesures prosodiques sur de larges bases de données acoustiques. L'utilisation de techniques d'apprentissage automatique consiste à réduire considérablement les temps d'acquisition des mesures de durée segmentale, du fondamental et des pauses sur des corpus de parole continue; cet ensemble de mesures définit une *base de mesures prosodiques*. Ce paragraphe présente tout d'abord le processus de création automatique des bases de mesures prosodiques puis la mise en oeuvre de ces principes pour la prédiction de la durée en Français et la prédiction du fondamental en Allemand.

4.1 Constitution de bases de mesures prosodiques

Pour la modélisation de la prosodie, un grand nombre de textes très variés a été enregistré par les locuteurs retenus pour l'enregistrement des répertoires d'unités acoustiques. Ces textes représentent pour chaque langue 2 heures de parole environ, soit approximativement 66000 segments acoustiques pour le Français. L'objectif consiste à déterminer pour chaque segment acoustique son étiquette phonétique, sa durée et les valeurs initiale et finale du fondamental. Chaque texte prononcé est découpé automatiquement en groupes de souffles séparés par des pauses acoustiques d'une durée supérieure à 1 seconde. Le module de transcription orthographique/phonétique du système de synthèse effectue la transcription phonétique de chaque phrase. Un découpage du signal acoustique en phonèmes est réalisé par l'alignement forcé d'une suite de modèles de Markov cachés sur la suite des trames acoustiques [Boëffard, 1993]. Il s'agit de modèles qui prennent en compte les principaux phénomènes d'assimilation et enrichissent ainsi la transcription phonétique initiale. Des informations de nature lexicale complètent la construction de modèles à variantes de prononciation; par exemple, des modèles de pause "possibles" sont insérés à chaque frontière de mot, de manière à pouvoir retrouver, après alignement, le découpage en pauses effectivement réalisé par le locuteur. Après segmentation, les signaux de parole sont soumis à une analyse fine du fondamental. Des erreurs sont bien évidemment présentes dans ces bases. Les messages synthétiques produits à partir de cette prosodie naturelle acquise automatiquement sont écoutés. Seuls les passages perceptuellement non acceptables sont éliminés de la base. En effet, l'utilisation ultérieure d'outils statistiques permet de faire l'hypothèse que des erreurs moins graves, si elles sont peu nombreuses, n'interviendront que de façon marginale dans la précision des modèles de prédiction de la prosodie. On obtient donc pour chaque phrase, son découpage en pauses, sa transcription phonétique effective (tenant compte des phénomènes d'assimilation), la durée et les valeurs initiale et finale du fondamental pour chacun des segments phonétiques voisés. Cette technique a été appliquée à la création des bases de mesures prosodiques en Français, Allemand, Anglais et Espagnol.

4.2 Durée segmentale en Français

Un premier essai d'apprentissage automatique de la durée des phonèmes a été mené en utilisant des arbres de classification [Riley, 1992]. Avec un tel système d'apprentissage, un vecteur de variables descriptives ou attributs est associé à la durée de chaque phonème. Les variables sont définies par un expert mais l'instanciation de ces variables a été acquise automatiquement à l'issue des traitements haut-niveaux de la synthèse du Français (CNETVOX [Larreur & al., 1989]). Un vecteur d'attributs contient un ensemble d'informations variées, telles que la nature du phonème, la syllabification du mot, la position du mot dans la phrase, le type de phrase, les marqueurs syntaxico-prosodiques ainsi qu'un certain nombre d'attributs contextuels. Une vingtaine d'attributs a été finalement retenue pour la prédiction de la durée en Français. Des techniques d'apprentissage d'arbre de classification comme l'algorithme CART [Breiman et al., 1984] ont été mises en oeuvres. A la synthèse, il suffit de présenter à l'entrée de l'arbre un vecteur d'attributs, de même composition que ceux définis précédemment, pour obtenir une valeur de durée pour chaque phonème à synthétiser. Une évaluation subjective comparant cette nouvelle approche au mo-

dèle de référence de CNETVOX a montré l'équivalence perceptuelle des deux modèles (test de préférence par paires AB/BA avec 8 auditeurs et 14 phrases avec le modèle de prosodie par apprentissage automatique versus le modèle de prosodie CNETVOX: score d'équivalence de 70%).

4.3 Contours mélodiques en Allemand

L'apprentissage de contours mélodiques a été mené, pour l'étude de l'Allemand, par l'utilisation de réseaux de neurones. La prédiction des contours mélodiques est effectuée à partir d'un étiquetage de la base de mesures prosodiques en groupes prosodiques et en accents de phrase. Dans une première expérimentation, les marqueurs syntaxico-prosodiques et la prédiction de l'accent sont issus du système de synthèse ALLVOC [Bigorgne & al., 1993]. Bien que ces marqueurs ne correspondent pas toujours au découpage prosodique et à l'accent effectivement réalisés par le locuteur, il est cependant possible d'utiliser cet étiquetage pour l'apprentissage d'un réseau de neurones. Les courbes d'intonation produites par ce système donnent une voix de synthèse de meilleure qualité que celle du système de référence de l'Allemand; des tests d'écoute sont en cours de réalisation. Les résultats d'un test informel d'écoute démontrent qu'il est déjà possible de produire une intonation de qualité en automatisant la construction de la base de mesures prosodiques et son étiquetage prosodique. Actuellement, l'optimisation manuelle réside dans la définition de la topologie du réseau de neurones employé. Par rapport à de précédents travaux sur l'apprentissage de contours mélodiques [Traber, 1992], le codage de l'entrée et de la sortie des réseaux a été modifié: une simplification du codage de l'entrée (ce qui mène à une réduction de l'espace des vecteurs d'entrée) et un meilleur codage de la sortie rendu possible par la segmentation en phonèmes et pas seulement en demi-syllabes.

5. Conclusion

Cet article a permis de présenter l'utilisation de techniques d'apprentissage automatique pour le développement des modules de traitements linguistique et prosodique des systèmes de synthèse multilingue du CNET. Ces techniques ont été appliquées au cours de l'élaboration récente du système de synthèse en Anglais. Elles ont été aussi mises en oeuvre sur des langues aux développements plus anciens comme le Français et l'Allemand.

6. Bibliographie

Bartkova, K., Sorin, C., 1987 A model of segmental duration for speech synthesis in French. Speech Communication, 6, North Holland, 245-260.

Bigorgne, D., Boëffard, O., Cherbonnel, B., Emerard, F., Larreur, D., Le Saint Milon, J.L., Métayer, I., Sorin, C., & White, S., 1993, "Multilingual PSOLA Test-to-Speech System", proceedings of the IEEE-ICASSP Conference, Vol.2, 187-190.

Brill, E., 1993, A Corpus-based approach to language learning; PhD Thesis, University of Pennsylvania.

Boëffard, O., 1993, "Segmentation automatique d'unités acoustiques pour la synthèse de la parole", Thèse de Doctorat, Université de Rennes I.

Breiman, L., Friedman, J.H., Olshen, R.A., & Stone, C.J., 1984, "Classification and regression trees", Belmont.

Conkie, A., & Bagshaw, P., 1995, "Traitements linguistiques et prosodiques pour la synthèse de la parole en Anglais", Note Technique, NT/LAA/TSS/613.

Emerard, F., Mortamet, L., & Cozannet, A., 1992, "Prosodic processing in a Text-To-Speech synthesis system using a database and learning procedures", Talking Machines, 225-247.

Hamon, C., Moulines, E., Charpentier, F., 1989, "A diphone synthesis system based on time-domain prosodic modifications of speech", Proceedings of the IEEE-ICASSP Conference, 238-241.

Hirschberg, J., & Prieto, P., 1994, "Training intonational phrasing rules automatically for english and spanish text-to-speech, Conference proceedings of the 2nd ESCA/IEEE Workshop on Speech Synthesis, 159-162.

Larreur, D., Emerard, F., Marty, F., 1989, "Linguistic and prosodic processing for a Text-To-Speech synthesis system", Proceedings of the Eurospeech Conference, Paris, 510-513.

Ostendorf, M., & Veilleux, N., 1994, "A Hierarchical Stochastic Model for Automatic Prediction of Prosodic Boundary location", Computational Linguistics, Vol. 20, N° 1, 27-54.

Riley, M.D., 1992, "Tree-based modelling of segmental durations", Talking Machines, 265-273.

Schnabel, B., 1988, "Développement d'un système de synthèse de l'Allemand à partir du texte", Thèse de Doctorat, Université Stendhal, Grenoble III.

Traber, C., 1992, "F0 generation with a database of natural f0 patterns and with a neural network", Talking Machines, 287-304.

JEP 96

RECONNAISSANCE DU LOCUTEUR

AVIGNON 10-14 JUIN 1996

ADAPTATION AU LOCUTEUR PAR CONVERSIONS SPECTRALES À L'AIDE DE RÉSEAUX NEURO-MIMÉTIQUES

Georges LINARES, Pascal NOCERA, Stephane IGOUNET

Laboratoire d'Informatique - 339, Chemin des Meinajariès - BP 1228 - 84911 Avignon Cedex 9
Tel: 90 84 35 20 - Fax: 90 84 35 01 - e-mail: linares,nocera,igounet@univ-avignon.fr

ABSTRACT

We present in this paper a connectionist method for reducing speech variability. This method is applied to speaker adaptation for automatic speech recognition systems.

This adaptation is based on an artificial neural network capturing a transformation function of a new speaker acoustic parameters into the reference speaker's.

Our system relies on a classic Markovian single-speaker speech recognition system. A large learning corpus is provided for a reference speaker to this system. In a recognition phase, artificial neural networks achieve the conversion towards the common référence. The result of which is decoded by the single-speaker recognition system. These conversions are achieved on the acoustic vectors by multi-layers backpropagation neural networks.

We have studied in particular the recognition rates by changing the size and the location of the temporal window at the networks inputs, and the size of the learning corpus provided for each new speaker.

Keywords:

Neural networks, speaker adaptation, Dynamic Time Warping, Markov models.

1. INTRODUCTION

Une des principales sources de difficulté du décodage de la parole est la variabilité du message oral. Cette variabilité peut avoir des origines très diverses : différences linguistiques, différences de réalisation acoustique du message, influence de l'environnement dans lequel le son est transmis, etc.. Les lois qui permettraient une normalisation des messages oraux ne sont pas toujours connues, et l'utilisation de techniques d'apprentissage automatique peut aider à résoudre ce type de problèmes. Nous proposons une méthode neuronale susceptible de rapprocher des réalisations acoustiques variables d'un message oral d'une réalisation de référence. Cette technique est ici appliquée à l'adaptation aux locuteurs d'un Système de Reconnaissance Automatique de la Parole (SRAP) : des réseaux de neurones réalisent les transformations des voix des nouveaux locuteurs en celle d'un locuteur de référence. Ces voix transformées sont ensuite décodées par un SRAP mono-locuteur.

Après avoir décrit le système de base, nous détaillerons notre méthode pour l'apprentissage des réseaux de neurones. Nos expérimentations portent sur l'influence du contexte acoustique local sur la qualité de la transformation apprise, puis sur la taille du corpus nécessaire à cet apprentissage.

2. LE SYSTÈME MONOLOCUTEUR

Afin de valider notre approche, nous avons construit, à l'aide de la boite à outils HTK [Young, 93], un système markovien de reconnaissance mono-locuteur. Ce système est uniquement destiné à évaluer la réduction de la variabilité et se limite à la reconnaissance des lettres de l'alphabet prononcées en mots isolés.

Dans notre système, les lettres sont divisées en 5 classes correspondant à 5 topologies différentes de type modèle de Bakis, qui déterminent les sources de Markov utilisées. Le locuteur de référence à prononcé 2115 lettres pour l'apprentissage et 1507 lettres pour le test. Le signal est paramétré par une suite de vecteurs à 26 coefficients calculés toutes les 12.5 ms dans des fenêtres de 25 ms. Ils sont composés des 12 coefficients MFCC, de l'énergie du signal, et des dérivées temporelles de ces 13 premiers paramètrès (calculées par différentiation des spectres successifs).

Nous obtenons 99.6% de décodage correct sur le locuteur de référence pour un modèle de lettres à 5 mixtures de gaussiènnes, et sans grammaire contrainte.

3. ARCHITECTURE DES RÉSEAUX DE NEURONES

Pour chaque locuteur testé, nous utilisons un perceptron multicouches qui produit en sortie un vecteur acoustique (26 noeuds). Les valeurs initiales des liens sont initialisées aléatoirement. Après quelques essais sur le nombre de couches et le nombre de nœuds par couche, nous avons retenu une architecture à une seule couche cachée contenant 15 nœuds..

La couche d'entrée sera constituée d'une séquence de vecteurs acoustiques destinée à modéliser le contexte du vecteur à convertir. L'influence du choix des séquences sur les résultats sera discutée plus loin (cf. *Couverture temporelle*).

4. ALIGNEMENT

Le rôle d'un réseau est de réaliser la transformation des vecteurs acoustiques d'un locuteur en ceux du locuteur de référence. L'apprentissage du réseau étant un apprentissage supervisé, il faut isoler pour chaque vecteur à convertir son correspondant parmi les vecteurs du locuteur de référence. Pour cela, nous utilisons une procédure d'alignement dynamique classique.

Cet alignement n'est utilisé que pour l'apprentissage des réseaux de neurones, et n'intervient pas lors de l'étape de reconnaissance.

Les distorsions temporelles à l'échelle du mot ne sont donc pas modélisées par le réseau de neurones, qui doit cependant apprendre les distorsions spectrales à l'échelle du vecteur acoustique. Cela signifie que les caractéristiques fréquentielles de la voix d'un nouveau locuteur seront transformées, mais que la dynamique de son élocution sera, elle, conservée.

5. COUVERTURE TEMPORELLE

Un système basé sur les modèles de Markov utilise l'évolution temporelle des vecteurs acoustiques. Or, bien que le réseau rapproche chaque vecteur de sa cible, leur transformation hors contexte peut altérer l'information contenue dans leur évolution. De plus, un même vecteur acoustique produit par un locuteur dans des contextes différents ne se traduira pas fatalement en un vecteur cible unique.

Extraire un vecteur de sa trajectoire comporte donc trois risques importants : pour une même donnée, proposer au réseau plusieurs conversions (ce qui perturberait son apprentissage), apprendre au réseau une conversion erronée ou très incomplète et enfin altérer l'évolution temporelle qui est utilisée au décodage.

Nous avons donc choisi de présenter des suites de vecteurs acoustiques au réseau. La difficulté est alors de connaître la largeur et la position de la fenêtre temporelle qui conditionne le vecteur à convertir. Nous avons testé plusieurs couvertures temporelles, en gardant à l'esprit que le véritable critère de réussite est l'amélioration des taux de reconnaissance du signal converti et non le rapprochement des vecteurs acoustique.

Pour évaluer l'influence de la fenêtre temporelle sur la qualité de la transformation spectrale, nous avons fournit pour chaque nouveau locuteur un corpus d'apprentissage constitué d'une seule fois l'alphabet. Les tests ont été réalisés sur un corpus constitué de 8 locuteurs masculins. Ils ont été obtenus au bout de 500 itérations de l'algorithme d'apprentissage. Nous avons défini un taux de rapprochement r entre le transformé T du vecteur souce S et sa cible C par :

$$r = 1 - \frac{\|T - C\|}{\|S - C\|}$$

La première colonne de la table 2 contient la liste des vecteurs acoustiques d'une séquence, désignés par leur position relative au vecteur courant (par exemple, "-1:0" représente le vecteur courant et son précédent). Pour chacune des séquences, sont présentés les taux de rapprochement sur le corpus d'apprentissage, puis les taux de rapprochement et de reconnaissance sur le corpus de test .

Table 1 : Evolution des taux de reconnaissance en fonction de la couverture temporelle

Couverture temporelle	Rapprochement (apprentissage)	Rapprochement (test)	Taux de reconnaissance (test)
0	48.8%	34.9%	75,96 %
-1:0	51.2%	35%	80,77 %
-2:-1:0	51	34.6%	82,69 %
-3:-2:1:0	50.6%	34.2%	75 %
0:1	50.5%	34.7%	78,85 %
0:1:2	51.2%	35.3%	**83.65** %
0:1:2:3	50.3%	35.3%	75 %
-1:0:1	50.6%	34.6%	**84,62** %
-1:0:1:2	51.2%	35.7%	80,77 %
-2:1:0:1:2	50.9%	34.7%	82,69 %
-5:0:1	50%	35.3%	82,69 %

On peut remarquer la faible corrélation des taux de rapprochement des paramètres avec la qualité du décodage, ce qui montre le peu de fiabilité de ce critère comme évaluateur de la qualité des conversions réalisées. On peut observer une dégradation des performances avec l'augmentation de la dimension de la couche d'entrée : l'élargissement des entrées implique une augmentation du nombre des exemples nécessaires à l'apprentissage.

6. TAILLE DU CORPUS D'APPRENTISSAGE

L'utilisation de cette méthode d'adaptation au locuteur ne se justifierait pas si la taille du corpus nécessaire au réseau était suffisante pour l'apprentissage d'un SRAP markovien. Il est donc important d'évaluer l'influence de la taille du corpus sur les performances de notre procédure d'adaptation.

Nous avons effectué plusieurs tests. Un premier essai a été réalisé en isolant 14 mots de l'alphabet où tous les phonèmes du vocabulaire apparaissent. L'apprentissage est fait sur ce corpus réduit, les tests sur l'ensemble du corpus de test (4 occurences de chacune des 26 mots de l'alphabet).

Les autres tests ont été effectués avec des corpus dans lesquels chaque lettre apparaît 1, 2, 4 et 8 fois, en utilisant une fenêtre temporelle composée du vecteur à convertir et de ses deux voisins (-1:0:1 avec la notation définie précédemment).

La figure 1 montre l'évolution de la distance relative des vecteurs transformés à leur cible et l'évolution des taux de reconnaissance en fonction de la taille des corpus d'apprentissage. La distance relative des transformés à leur cible est le rapport de la distance après transformation sur la distance avant transformation, exprimé en pourcentage.

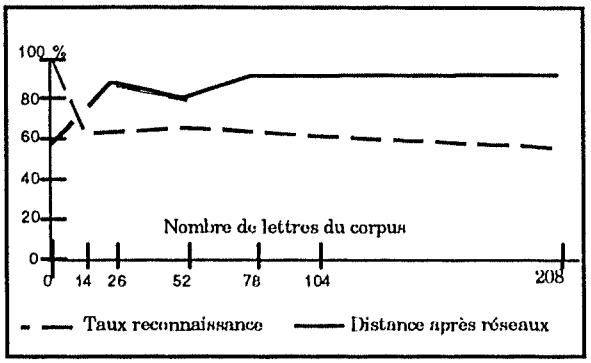


Figure 1 : Taux de reconnaissance en fonction de la taille du corpus

Présenter deux fois chaque lettre semble moins efficace que ne les présenter qu'une seule fois. Il est probable que, lorsque une lettre apparaît deux fois dans le corpus, on puisse associer à une même entrée deux cibles différentes, ce qui perturbe l'apprentissage du réseau. Ce risque est moindre en ne présentant qu'une fois chaque mot, tandis que l'utilisation d'une masse de données importante permet au réseau de réduire l'influence d'une telle incohérence du corpus. On peut cependant observer le peu d'écart entre les résultats obtenus avec le corpus de 26 lettres et les corpus de tailles supérieures.

7. RÉSULTATS

Afin d'estimer l'amélioration des performances apportée par l'usage de notre procédure d'adaptation, nous avons fait un premier décodage des données de test sans traitement par les réseaux neuro-mimétiques. Nous obtenons une moyenne de 64% de décodage correct. Les scores varient de 24% à 80,7%, ce qui représente des écarts très importants.

Nous avons ensuite utilisé notre procédure d'adaptation, avec la couverture temporelle (-1:0:1) et un corpus d'apprentissage constitué d'une seule occurence des 26 mots du vocabulaire.

Nous obtenons une moyenne de 88% de décodage correct. Les scores des 8 locuteurs varient de 78,85% à 96,3%.

8. CONCLUSION ET PERSPECTIVES

Nos résultats montrent que la capture par le réseau neuro-mimétique d'une fonction de conversion d'une voix en une autre ne permet qu'une imitation incomplète du locuteur cible, les distances entre vecteurs acoustiques n'étant réduites que d'environ 40%. Mais ce rapprochement des paramètres ne suffit pas à évaluer l'amélioration du décodage apportée par l'utilisation du réseau de neurones, et le rapprochement des voix des nouveaux locuteurs de celle du locuteur de référence permet d'améliorer très sensiblement les taux de reconnaissance pour l'ensemble des locuteurs. Ceci met en évidence l'amélioration qualitative de la transformation apportée par la prise en compte du contexte.

Par ailleurs, nous avons observé que l'augmentation de la taille du corpus d'apprentissage n'apportait qu'un gain limité de performances. L'apprentissage par un réseau de neurones d'une fonction aussi complexe que la transformation spectrale d'une voix en une autre semble donc nécessiter relativement peu de données d'apprentissage.

La méthode utilisée ici pour la réduction de la variabilité inter-locuteurs pourrait être appliquée à la réduction d'autres types de variabilités. En particulier, nous envisageons son utilisation dans le traitement de la voix des scaphandriers en grandes profondeurs. En effet; l'usage de mélange gazeux pour la respiration hyperbare transforme la voix des plongeurs. L'usage de réseaux neuronaux pour la réduction de cette variabilité due à l'environnement pourrait permettre d'améliorer l'intelligibilité des messages.

9. BIBLIOGRAPHIE ET RÉFÉRENCES

Bourlard H., Wellekens C.J. (1987) Multilayer Perceptrons and Automatic Speech Recognition, *IEEE First Annual International Conference on Neural Networks,* San Diego

Choukri K. (1987) Quelques approches pour l'adaptation au locuteur en reconnaissance automatique de la parole, *Thèse ENST*

Choukri, Montacié C., Chollet G.(1990) Transformations spectrales linéaires et neuro-mimétiques pour l'adaptation au locuteur, actes du séminaire *Variabilité et spécificité du locuteur : Etudes et Applications,* Marseille Luminy

Nocera P., Bulot P. (1990) Rule driven neural networks for speech recognition, *INNC 90* (IEEE), Paris, 31.5 pp 1353-1356.

Nocera P. (1992) Utilisation conjointe de réseaux neuronaux et de connaissances explicites pour le décodage acoustico-phonétique, *Thèse de l'Université d'Avignon*

Rumelhart D.E, Hinton G.E, Williams R. (1986) Learning internal representations by Back-Propagation, dans *Parallel Distributed Processing: Explorations in the microstructure of cognition,* MIT Press

Young S.J., Woodland P.C., Byrne W.J. (1993) HTK : Hidden Markov Model Toolkit V1.5 : Reference Manual, Cambridge University Engineering Department Speech Group and Entropic Research Laboratories Inc.

Amélioration des performances de vérification du locuteur par combinaison de méthodes

Dominique Genoud* Guillaume Gravier* Frédéric Bimbot++
Gérard Chollet*++

* IDIAP, CP 592, CH-1920 Martigny, Suisse.
email: genoud@idiap.ch

++CNRS URA820 ENST Rue Louis Barrault 46,
Paris France. email:
chollet@sig.enst.fr,bimbot@sig.enst.fr

1 Abstract

Moving from laboratory to real applications is often, for speaker verification systems, a very disappointing experience. Among the known problems, the lack of training speech data is a crucial one. Even if very powerful algorithms are available now, problems with a priori thresholds setting due to the amount of training data can decrease drastically the performances. The aim of this paper is to describe how the combination of algorithms with a priori decision thresholds can improve the overall robustness of a speaker verification application. The evaluation is performed in the context of a field application where each client is verified from a 7 digit pin code.

2 Introduction

Le transfert de systèmes de vérification du locuteur du laboratoire à des applications réelles peut être une expérience très décevante, surtout si les données disponibles sont de la parole téléphonique. Parmi les problèmes souvent rencontrés, citons le manque de données à disposition pour construire des références correctes du client inscrit. Ce manque de données entraîne souvent une baisse de performances importante de méthodes pourtant efficaces sinon. Un deuxième problème souvent rencontré lors d'applications réelles est le comportement très différent d'une méthode selon les locuteurs. Ce papier montre qu'il est possible d'améliorer les performances globales du système de vérification en combinant entre eux les résultats de plusieurs méthodes.

Nous allons, après avoir décrit l'application et les méthodes utilisées, analyser l'effet de la détermination de seuils a priori, de l'amélioration de ceux-ci, pour finalement nous pencher sur la combinaison de décisions permettant d'améliorer le résultat final.

3 Description de l'application

L'application consiste à sécuriser l'accès à un serveur d'information téléphonique personnalisé. Le système est basé sur une liaison téléphonique RNIS. Cette application comporte 2 phases : l'inscription et l'accès au service.

Lors de l'inscription, le locuteur prononce son nom, prénom, adresse, tous les chiffres de 0 à 9 en séquence, et 5 fois son code client composé de 7 chiffres.

Lors de l'accès, le locuteur prononce une fois son code client, celui-ci est ensuite vérifié par le système. La vérification se déroule en deux étapes: (1) tout d'abord la séquence de chiffres est reconnue en utilisant un système de reconnaissance de la parole indépendante du locuteur basé sur une technique HMM [Gro93]. (2) La séquence de parole est ensuite comparée aux références du locuteur correspondant au code client reconnu durant la première phase. La comparaison aux références permet de prendre la décision d'accepter ou de rejeter le locuteur. De plus si la comparaison ne peut être effectuée de façon assez fiable, le système peut décider de douter. En cas de doute une question est posée au locuteur une vérification indépendante du texte est effectuée, aboutissant au rejet ou à l'acceptation de l'appelant.

Dans cet article nous ne considérerons que de la partie de vérification dépendante du texte utilisant le code client prononcé par le locuteur.

4 Combinaison de méthodes

Dans le problème qui nous préoccupe ici, la combinaison de méthodes vise à obtenir, par l'utilisation de différentes mesures, une meilleure information

globale sur le locuteur [Ant95]. En effet, si chacune de ces mesures ou méthodes donnent une image impropre mais différente du locuteur, leur combinaison devrait améliorer le résultat final. Dans le cas de la vérification du locuteur, la réponse finale du système est très simple: soit l'on accepte le morceau de parole testé comme appartenant au locuteur proclamé, soit on le rejette. Cette prise de décision peut se faire à différents niveaux [Das94]. Soit, par exemple, chaque méthode prend une décision partielle et on combine ces décisions pour prendre une décision finale, soit on combine les mesures fournies par chaque méthode et on prend une décision sur le résultat de la fusion de ces mesures. Pour cet article nous avons choisi la première approche pour la combinaison de méthodes.

5 Les méthodes de vérification

Trois méthodes de vérification dépendantes du texte ont été utilisées. Ces trois méthodes acceptent en entrée un ensemble de vecteurs de 12 coefficients cepstraux LPC .

5.1 Dynamic Time Warpping (DTW)

L'algorithme de DTW est utilisé depuis fort longtemps en reconnaissance de la parole, il est aussi utilisé avec succès en vérification du locuteur. Principalement il consiste à effectuer une comparaison dynamique entre une matrice de référence et une matrice de test. Le résultat est une mesure de distance entre le test et la référence.

5.2 Statistique du second ordre (SSO)

Dans cet algorithme une matrice de covariance $X = \frac{1}{M}\sum_{t=1}^{t=M} X_t X_t^T$ de la séquence de parole de référence est crée.

Une matrice de covariance Y de la séquence de test est aussi crée. On extrait ensuite les valeurs propres λ_i de YX^{-1}.

Une mesure de sphéricité symétrique $\mu_{AH}(X,Y)$ est ensuite effectuée [BM94].

- $\mu_{AH}(X,Y) = \log[\frac{\mathbf{A}}{\mathbf{H}}]$
- $\mathbf{A}(\lambda_1, \lambda_2, \ldots, \lambda_m) = \frac{1}{m}\sum_{i=1}^{i=m} \lambda_i$
- $\mathbf{H}(\lambda_1, \lambda_2, \ldots, \lambda_m) = m(\sum_{i=1}^{i=m} \frac{1}{\lambda_i})^{-1}$

en utilisant la relation $\mu_{AH}(X,Y) = 0$ <==> A = H <==> 'X proportionnel à Y, on peut déterminer une distance entre les séquences de référence et de test.

5.3 Modèles de Markov cachés (HMM)

Deux sortes de modèles HMM [Gok91] sont créés pour chaque chiffre de 0 à 9: (1) *Un modèle du monde*, créé a partir d'une base de donnée (Polyphone [Lan95]) comportant un grand nombre de locuteurs dont on a extrait 300 répétitions de chaque chiffre. Les paramètres de ce modèle sont estimés par un entraînement standard (Initialisation par l'algorithme de Viterbi, Réestimation de paramètres par Baum-Welch) [Gro93] ce modèle est identique pour tous les clients. (2) *Un modèle du client*, qui utilise comme paramètres initiaux le modèle du monde, et dont on réestime les paramètres pour chaque locuteur avec les données de celui-ci, ce modèle est donc dépendant du client. Lors de l'accès, pour chaque chiffre du code que prononce le locuteur, on calcule le rapport de vraisemblance

$L_{mc} = \log(V_c) - \log(V_m)$.

avec V_c, V_m les vraisemblances obtenues pour le modèle du client et le modèle du monde respectivement.

Les modèles ont tous la même structure HMM gauche-droite constituée d'un état par phonème et un état par transition entre phonème [Gra95].

6 Détermination des seuils

Pour qu'une décision d'acceptation ou de rejet d'un locuteur puisse être prise il est nécessaire de fixer un seuil de décision avant d'effectuer la vérification (seuil a priori). Nous avons choisi 3 méthodes de réglage du seuil. Chacune essayant, pour chaque méthode, de modéliser de façon différente la séparation des scores obtenus par les locuteurs ou les imposteurs.

La détermination de certains paramètres nécessaire au réglage de ces seuils s'est faite sur une base de données d'imposture (Polycode [GC95]). Cette base de données est constituée de 25 locuteurs prononçant des séquences de 10 chiffres et des séquences de 7 chiffres. Des séquences de 10 chiffres sont extraits les chiffres du code client dont on veut faire l'imposture, le code étant recrée par concaténation des chiffres extraits. Le paragraphe "Résultats" donne plus de détails sur la répartition des données de la base utilisée. Toutes les séquences utilisées pour le réglage des seuils sont *différentes* des données utilisées pour

l'entraînement et le test.

6.1 Seuil EER global

Ce seuil est déterminé, pour chaque méthode, par l'EER (Equal Error Rate) des 25 locuteurs d'une base de données de développement, différente des données d'entraînement et de test. Ce seuil est donc fixé a priori avant le test.

6.2 Seuil EER individuel

Ce seuil est déterminé au moment de l'entraînement avec l'EER sur les séquences prononcées par le locuteur et par les séquences d'imposture de 5 autres locuteurs.

6.3 Seuil individuel par imposture (FURUI)

Furui a montré [FUR94] que lorsque très peu de données d'entraînement sont à disposition, il est intéressant de fixer le seuil de décision uniquement par les tentatives d'imposture. La fixation du seuil se déroule en deux phases: (1) Deux constantes $C1, C2$, sont estimées sur la base de données des 25 locuteurs par régression linéaire. Ces constantes sont valables pour tous les locuteurs. (2) Lors de l'inscription le seuil du locuteur est estimé en utilisant

$Seuil_x = C1(\mu_x - \sigma_x) + C2$

où μ_x, σ_x sont les paramètres de la gaussienne $N(\mu, \sigma)$ estimée sur les scores d'imposture.

7 Combinaison des décisions

Afin d'améliorer la décision globale de notre système nous avons combiné les décisions de chaque méthode (DTW, SSO, HMM). La combinaison utilisée ici est un système majoritaire pondéré avec seuil de doute. D'autre types de combinaisons de méthodes ont été utilisés [The93, Ant95, Das94].

Chaque méthode prend une décision d'acceptation ou de rejet par rapport au seuil fixé a priori. Cette décision est ensuite pondérée (entre 0 et 1) selon la distance à laquelle on se trouve du seuil, ce qui revient à donner une confiance dans cette décision.

Le système compare ensuite la moyenne des confiances des méthodes qui ont obtenu la majorité à un seuil (seuil de doute). Si cette moyenne est supérieure au seuil on prend la décision de la majorité sinon il y a doute.

8 Résultats

La base de données Polycode [GC95] qui a servi à nos expériences est constituée de 25 locuteurs prononçant des séquences de 10 chiffres et de 7 chiffres. Les 25 locuteurs prononçant chacun 5 séquences de 10 chiffres et 5 séquences de 7 chiffres ont été utilisés pour la détermination des seuils EER globaux et des constantes $C1, C2$ de la méthode Furui. Nous avons extrait 10 Locuteurs de Polycode pour effectuer nos expérimentations. Les données d'entraînement sont prises sur *une* session d'enregistrement. Les données de test sont prises sur plusieurs sessions d'enregistrement. 200 tests d'accès sont effectués pour chaque locuteur. Sur ces 200 tests, 20 sont des vrais accès du locuteur et 180 des accès d'imposteurs. Pour les impostures on utilise le même procédé de concaténation que celui cité au paragraphe 4. Les données utilisées pour le calcul des seuils, pour l'entraînement et pour les tentatives d'accès sont toutes différentes.

Méthode	FR%	FA%	NB Tests
DTW	100	0.0	2000
SOSM	51.0	0.0	2000
HMM L/R	28.5	1.05	2000
Décision Combinée	33.5	3.05	2000

Table 1: Performance des méthodes avec seuil EER individuel

La table 1 montre, pour le cas d'un seuil EER individuel, à quel point la fixation du seuil se révèle mauvaise avec le peu de données d'apprentissage dont nous disposons. La combinaison des méthodes dans ce cas n'offre pas de meilleures performances que la meilleure des méthodes.

Méthode	FR%	FA%	NB Tests
DTW	31.5	21.78	2000
SOSM	22.2	30.3	2000
HMM L/R	2.5	5.33	2000
Décision Combinée	8.0	3.17	2000

Table 2: Performance des méthodes avec seuil EER global

La table 2 montre qu'en utilisant un seuil EER global, la méthode HMM, dû à la normalisation par le monde que nous effectuons, se comporte mieux que les autres méthodes. La combinaison des méthodes ne donne pas non plus dans ce cas de résultats intéressants.

Avec l'utilisation d'un seuil déterminé par la méthode de Furui, la table 3 montre que , non seulement chaque méthode devient plus performante, mais la combinaison de celles-ci donne améliore cette fois la performance globale du système.

La table 4 nous indique le comportement des performances du système lors du changement de seuil de doute, ce seuil permet de régler ainsi les taux de faux rejets et fausses acceptations selon le niveau de sécurité requis par l'application.

Méthode	FR%	FA%	NB Tests
DTW	23.5	7.67	2000
SOSM	14.0	5.28	2000
HMM L/R	5.53	2.72	2000
Décision Combinée	2.0	2.72	2000

Table 3: Performance des méthodes,seuil FURUI

Seuil de doute	FR%	FA%	Doute%
0.2	2.0	2.72	0.0
0.5	2.0	2.33	0.83
0.7	2.0	1.11	10.11
0.8	1.0	0.89	20.2

Table 4: Performance lors de changement de seuil de doute (2000 tests)

9 Conclusion

Nous avons pu montrer que les performances générales d'un système de vérification du locuteur pouvaient être améliorées en utilisant une combinaison de méthodes avec un choix de seuil approprié. La détermination du seuil par la méthode de Furui confirme sa valeur en présence de peu de données d'entraînement. La combinaison de méthodes semble prometteuse, et il conviendrait de poursuivre ces expériences, soit en combinant non plus des décisions, mais des mesures, soit en ajoutant plus de méthodes, par exemple des méthodes basées sur des réseaux de neurones, ou utilisant une paramétrisation du signal différente cela afin de parcourir l'espace des mesures de la parole du locuteur le mieux possible.

10 Remerciements

Nous remercions Mehdi Homayounpour pour la mise à disposition de ses logiciels de DTW et de la détermination des seuils par la méthode FURUI.

Nous remercions également les Telecom PTT Suisses pour leur soutien dans le cadre des projets ATTACKS et AVIS. Nous remercions également la Communauté Européenne pour son soutien dans le projet Telematics CAVE (CAller VErification).

Bibliographie

[Ant95] Richard T. Antony. *Principles of Data Fusion Automation.* Artech House, 685 Canton Street Norwood,MA 02062, 1995.

[BM94] Frédéric BIMBOT and Luc MATHAN. Second-order statistical measures for text-independant speaker identification. In ESCA [ESC94], pages 51–54.

[Das94] Belur V. Dasarathy. *Decision Fusion.* IEEE Computer Society Press, Los Alamitos, California, 1994.

[ESC94] ESCA, editor. *ESCA Workshop on Automatic Speaker Recognition Identification Verification.* ESCA, April 1994.

[FUR94] Sadaoki FURUI. An overview of speaker recognition technology. In ESCA [ESC94], pages 1–9.

[GC95] Dominique Genoud and Gérard Chollet. Polycode a verification database. Technical report, IDIAP, CH-1920 Martigny, 1995.

[Gok91] A.E. Rosenberg & C.H. Lee & S. Gokoen. Connected word talker verification using whole word hidden markov model. In *ICASSP-91*, pages 381–384, 1991.

[Gra95] Guillaume Gravier. Vérification du locuteur par modèles de markov cachés gauche-droite. Rapport de stage dea, IDIAP, CH-1920 Martigny, 1995.

[Gro93] Cambridge University Speech Group. *HTK Hidden Markov Model Toolkit.* Entropic Research Laboratories Inc., Cambridge, December 1993.

[Lan95] G. Chollet & J.L. Cochard & A. Constantinescu & Ph. Langlais. Swiss french polyphone and polyvar: telephone speech databases to study intra and inter speker variability. Technical report, IDIAP, 1995.

[The93] Ph. Thevenaz. *Résidu de prédiction linéaire et reconnaissance de locuteurs indépendante du texte.* PhD thesis, Université de Neuchâtel, 1993.

COOPERATION ET COMPETITION DE MODELES EN RECONNAISSANCE DU LOCUTEUR

J.-L. Le Floch, C. Montacié & M.-J. Caraty

LAFORIA-IBP, Université Paris 6, CNRS-URA 1095
4 place Jussieu, 75252 Paris Cedex 5, FRANCE

ABSTRACT

In order to improve the performances of speaker recognition on telephone speech, we present methods based on the utilisation of several AR-vector models per speaker and a cooperation of AR-vector model and GMM.

These approaches allow us to obtain the best known performances on the NTIMIT database: 81.4 % on 168 speakers and 65.0 % on 630 speakers.

1. INTRODUCTION

La reconnaissance du locuteur fait référence à deux problèmes qui sont la vérification du locuteur et l'identification du locuteur. Dans cet article, nous nous sommes particulièrement intéressés à l'identification du locuteur sur un signal de qualité téléphonique.

Nous avons comparé les résultats de trois systèmes de reconnaissance du locuteur sur les bases de données TIMIT et NTIMIT et observé une baisse importante des performances dans le cas du signal de qualité téléphonique (NTIMIT). Dans le but de résoudre ce problème, nous présentons des méthodes basées sur l'utilisation de plusieurs modèles AR-vectoriels par locuteur: une coopération et une compétition de modèles appris sur des segments phonétiques distincts. Ces méthodes pouvant s'étendre à l'utilisation de modèles de natures différentes, nous présentons une coopération des modèles AR-vectoriels et des GMM.

2. DEGRADATION DES RESULTATS SUR LA BASE DE DONNEES NTIMIT

Plusieurs systèmes d'identification du locuteur existants obtiennent de bonnes performances sur de la parole non-bruitée. Nous avons choisi de tester la difficulté de l'identification du locuteur sur les trois systèmes suivants représentant l'état de l'art:

- Modèle de Mélange de Gaussiennes (GMM).
- Mesure Statistique du Second Ordre (MSSO).
- Modèle AR-Vectoriel (MARV).

Dans le GMM (Reynolds, 1994), la distribution des vecteurs spectraux est modélisée par une somme pondérée de densités de probabilités gaussiennes.

La MSSO (Bimbot & al., 1995) est une mesure sur les matrices de covariance des vecteurs spectraux.

Le MARV (Montacié & al., 1992, 1996) est un modèle de prédiction linéaire des vecteurs spectraux de taille m à partir des p vecteurs précédents.

Les bases de données TIMIT et NTIMIT (Fisher, 1986) sont composées de dix phrases par locuteur. Huit phrases sont utilisées pour l'apprentissage des modèles, les deux restantes sont utilisées séparément pour le test. Ces conditions expérimentales sont celles des travaux de (Reynolds, 1994 et 1995) sur TIMIT et NTIMIT. Nous avons implémenté la MSSO et les MARV pour rester dans les mêmes conditions expérimentales.

Un débruitage est réalisé sur chaque phrase avant l'estimation des modèles. Ce prétraitement consiste en un filtrage dans la bande téléphonique suivi d'une segmentation bruit/parole basée sur l'énergie du signal.

Table 1: Résultats en identification du locuteur pour les systèmes GMM, MSSO et MARV.

	TIMIT		NTIMIT	
nb de locuteur	168	630	168	630
GMM (Reynolds, 1994 et 1995)	99,7	99,5	76,2	60,7
MSSO	97,6	96,7	60,1	43,0
MARV (p=2, m=12)	99,0	98,2	69,4	51,8
MARV (p=3, m=12)	------	------	72,6	58,0
MARV (p=3, m=14)	------	------	78,1	64,6

Le résultat des tests montre une forte dégradation pour chaque système sur la base de données NTIMIT (table 1). Les chutes des pourcentages de reconnaissance de l'ordre de 20 % à 30 % démontrent la difficulté de la tâche ainsi que la nécessité de développer des systèmes plus performants et de comprendre les faiblesses et les capacités d'un système de reconnaissance du locuteur.

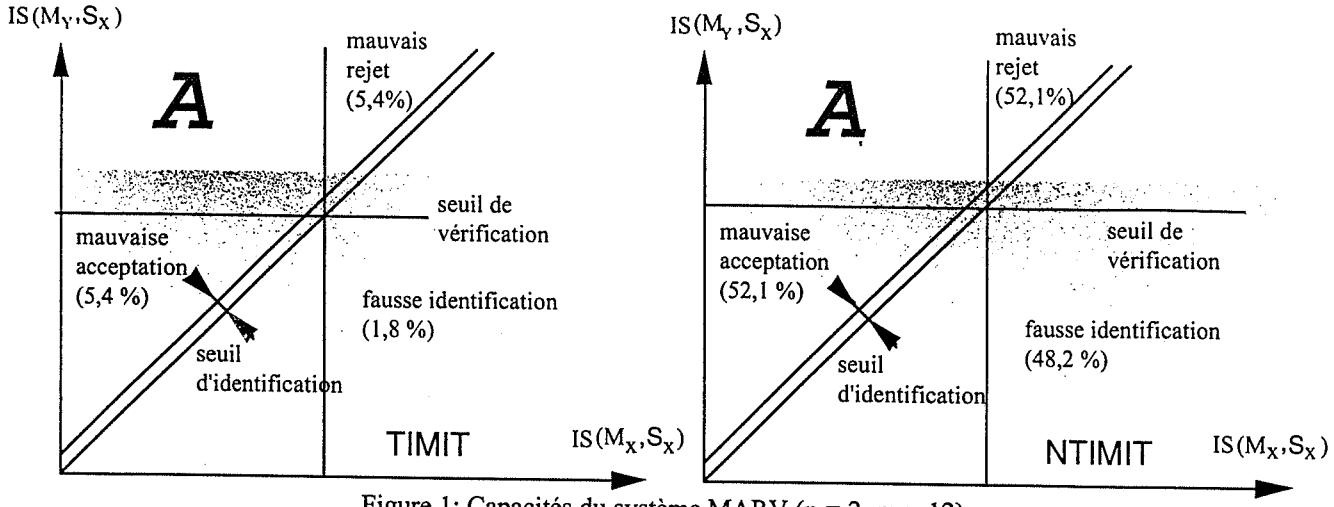


Figure 1: Capacités du système MARV (p = 2, m = 12).

3. CAPACITES D'UN SYSTEME DE RECONNAISSANCE DU LOCUTEUR

Les taux d'identification ne donnent pas assez d'information sur les capacités d'un système à reconnaître les locuteurs. L'illustration graphique (figure 1) de la comparaison entre les mesures inter et intra-locuteur est plus informative.

Soient:

S_X une phrase test du locuteur X,
M_X le modèle de référence du locuteur X,
M_Y le modèle de référence le plus proche de S_X et différent du modèle M_X.

L'abscisse représente la mesure IS (Montacié & al., 1992) entre S_X et M_X. L'ordonnée représente la mesure entre S_X et M_Y. Chaque point peut être expliqué de deux points de vue, le premier en termes d'identification du locuteur, le second en termes de vérification du locuteur. En premier lieu, si l'abscisse est plus grande que l'ordonnée, le vrai locuteur n'est pas identifié: il y a une fausse identification. Du second point de vue, l'abscisse représente un test de vérification où le locuteur X n'est pas un imposteur. Si cette valeur est plus grande que le seuil de vérification, il y a rejet du locuteur X. L'ordonnée représente un test de vérification où le locuteur X se fait passer pour le locuteur Y. Si cette valeur est inférieure au seuil de vérification, il y a acceptation de l'imposteur.

La distribution et la localisation des points permettent l'estimation des résultats de reconnaissance quelques soient les modifications des seuils de vérification et d'identification. Pour un système à hautes performances, les points doivent se trouver dans la zone A. Si les tests des MARV sur TIMIT sont bien localisés dans la zone A, on observe un glissement pour NTIMIT imputable à la faible résistance du système aux conditions téléphoniques.

4. SELECTION DE CLASSES PHONETIQUES POUR LA RECONNAISSANCE DU LOCUTEUR

Un système de vérification du locuteur peut être trompé par un imposteur possédant l'enregistrement d'un des locuteurs autorisés. Pour éviter ce type d'imposture, le système doit imposer la lecture d'une phrase choisie aléatoirement et vérifier le texte prononcé. Le système réalise dans le même temps une segmentation phonétique grossière et peut donc utiliser le fait que certains segments phonétiques contiennent plus d'informations caractéristiques du locuteur que d'autres. Dans une étude précédente (Le Floch & al., 1994), les meilleurs résultats ont été obtenus sur les voyelles, les diphtongues, les nasales et les liquides/glissées. Huit classes phonétiques sont définies (table 2).

Table 2: Description des classes phonétiques à partir des symboles phonétiques utilisés pour la transcription des phrases de la base de données TIMIT.

1	Voyelles
2	Diphtongues
3	Nasales
4	Liquides, Glissées
5	Voyelles, Diphtongues
6	Nasales, Liquides, Glissées
7	Voyelles, Nasales, Diphtongues, Liquides, Glissées
8	Voisées

Pour tirer parti de différents modèles AR-vectoriels par locuteur, nous avons étudié deux approches. l'approche segmentale a pour but de déterminer si des modèles AR-vectoriels appris sur des segments phonétiques choisis permettront une amélioration des performances. L'approche analytique a pour but la coopération ou la compétition de plusieurs modèles AR-vectoriels appris sur différentes classes phonétiques.

4.1. Approche Segmentale

Dans l'approche segmentale un modèle AR-vectoriel est appris par locuteur sur la concaténation de segments phonétiques sélectionnés. Seuls les segments sélectionnés sont pris en compte au cours de l'apprentissage et du test. Toutes les expériences sont réalisées sur 168 locuteurs de la base de données NTIMIT avec des modèles AR-vectoriels d'ordre 3 et des vecteurs spectraux d'ordre 14.

Table 3: Pourcentages d'identification du locuteur sur les classes phonétiques choisies et proportion des classes phonétiques dans les phrases.

Classes phonétiques	1	2	3	4
Proportion de parole	33,3	8,3	7,2	8,2
Taux d'identification	66,8	23,1	28,1	15,3
Classes phonétiques	5	6	7	8
Proportion de parole	41,6	15,4	57,0	64,0
Taux d'identification	76,9	35,9	80,2	79,6

En utilisant une partie de la phrase (i.e., 57 % de la phrase, classe {7}), nous obtenons un meilleur résultat qu'en utilisant la phrase entière (78,1 % vs 80,2 %). La même expérience sur la base complète (630 locuteurs) ne permet cependant qu'une faible augmentation du résultat (65,0 % vs 64,6 %). Plus aisément détectables en conditions réelles, les voisées {8} atteignent un résultat similaire à celui de la classe {7} (79,6 % vs 80,2 %).

L'augmentation des performances est vraisemblablement due à la sélection de phonèmes contenant des informations discriminant les locuteurs. Ces informations pouvant être de natures différentes suivant le type de phonèmes, un traitement global ne peut les utiliser pleinement. Nous allons développer dans l'approche analytique des méthodes utilisant des modèles appris sur différentes classes phonétiques pour tenter d'extraire le maximum d'information de ces classes.

4.2. Approche analytique

Dans l'approche analytique, en coopération et en compétition de modèles, la décision finale d'identification du locuteur est prise à partir de l'ensemble des tests d'identification locale effectuée pour chaque vecteur spectral.

Définissons les termes suivants:

N: nombre de vecteurs,
N_L: nombre de locuteurs,
N_M: nombre de modèles par locuteur appris sur différentes classes phonétiques.

Après analyse, une phrase de test consiste en une suite de vecteurs spectraux. Pour chaque vecteur, $N_M * N_L$ erreurs de prédiction sont calculées à partir des N_M modèles de chacun des N_L locuteurs. Il en résulte N_M tests d'identification locale par vecteur spectral ce qui corespond à un test par classe phonétique.

4.2.1 Coopération de modèles

A partir des tests d'identification locale effectués pour chaque vecteur spectral et chacune des N_M classes phonétiques, nous définissons la décision finale $D^{\{\omega_j\}}$ comme la formule:

$$D^{\{\omega_j\}} = \underset{L=(1,\ldots,N_L)}{\text{Argmax}} \left[\sum_{k=1}^{N_M} \sum_{i=1}^{N} \sum_{j=1}^{N_L} \omega_{ij} \, O(L_{ijk}, L) \right]$$

avec $\{\omega_{ij}\}$: $(j = 1,..,N_L)$ ensemble des pondérations pour le vecteur v_i,
L_{ijk}: $j^{ième}$ plus proche locuteur du vecteur v_i pour la classe phonétique k (L_{ijk} est issus des tests d'identification locale)
$O(L_{ijk}, L) = 1$ si $L_{ijk} = L$, 0 sinon.

Nous avons essayé cinq ensembles de pondérations $\{\omega_{ij}\}$ communément utilisés (Dudani, 1976). L'ensemble permettant d'obtenir les meilleurs résultats est le suivant:

$$\{\omega_{ij} = \frac{1}{j}\} \; (j = 1, \ldots, N_L), \; (i = 1, \ldots, N).$$

Mais ce type de pondérations (Le Floch & al., 1995) ne prend pas en considération les différences de proximité entre les locuteurs. L'ensemble de pondérations suivant résout ce problème:

$$\{\omega_{ij} = \frac{E_{i2k}}{E_{i1k}}\} \; j = 1, \; (i = 1, \ldots, N).$$

$$\{\omega_{ij} = \frac{E_{i1k}}{E_{ijk}}\} \; \begin{matrix} (i = 1, \ldots, N) \\ (j = 2, \ldots, N_L) \end{matrix}$$

où E_{ijk} est l'erreur de prédiction sur le vecteur v_i du $j^{ième}$ plus proche locuteur pour la classe phonétique k.

Table 5: Taux d'identification du locuteur au niveau analytique pour les classes phonétiques {5} et {6} ainsi que pour la coopération des classes phonétiques {5} et {6}.

Classe phonétique	5	6	5&6
Taux d'identification	74,3	39,2	79,3

Les résultats en coopération de modèles de l'approche analytique sont meilleurs que ceux de l'approche segmentale (79,3 % vs 76,9 %) mais restent légèrement inférieurs à ceux de l'approche segmentale globale (79,3 % vs 80,2 %). La coopération de modèles impose le modèle utilisé pour la mesure sur un vecteur. Une compétition de modèles devrait permettre par la seule prise en compte des meilleurs modèles d'améliorer les performances.

4.2.2 Compétition de modèles

Dans cette approche nous mettons en concurrence plusieurs modèles AR-vectoriels. La mesure locale entre le locuteur L et le vecteur v_i est la suivante:

$$Ml(L,v_i) = \sum_{k=1}^{N_M} \frac{1}{k} E_{iLk}$$

où E_{iLk} est l'erreur de prédiction sur le vecteur v_i du $k^{ième}$ meilleur modèle du locuteur L.

La décision finale est réalisé de la manière suivante:

$$D = \underset{L=(1,...,N_L)}{Argmin} \left[\sum_{i=1}^{N} Ml(L,v_i) \right]$$

avec $Ml(L,v_i)$: mesure locale entre le locuteur L et le vecteur v_i.

Les classes phonétiques utilisées pour la concurrence de modèles sont les suivantes: {1}, {2}, {3}, {4}, {5}, {8}, les non-voisées et l'ensemble des phonèmes.

Table 6: Taux d'identification du modèle ARV global, du modèle ARV calculé sur la classe phonétique {7} et de la compétition de modèle au niveau analytique.

MARV phrase entière	78,1 %
MARV classe phonétique {7}	80,2 %
compétition de modèle ARV	81,4 %

La compétition de modèles au niveau analytique donne un résultat très prometteur. Mais nous n'avons abordé que des coopérations ou compétitions de modèles de même nature. Une coopération de deux modèles de natures différentes peut permettre d'éviter certaines erreurs qui ne sont pas communes aux deux modélisations.

5. COOPÉRATION DES MARV ET DES GMM

La coopération de deux méthodes de natures différentes n'est possible qu'après une normalisation de leur mesures. La décision est réalisé de la manière suivant:

$$D = \underset{L=(1,...,N_L)}{Argmin} \left[Mes_{ARV}(L) + Mes_{GMM}(L) \right]$$

où $Mes_{ARV}(L)$: mesure normalisée sur la phrase de test à partir du modèle AR-vectoriel du locuteur L.

$Mes_{GMM}(L)$: mesure normalisée sur la phrase de test à partir du GMM du locuteur L.

Pour cette expérience, nous avons du implémenter les GMM, ce qui explique la différence de résultats avec ceux de Reynolds.

Table 7: Taux d'identification du modèles AR-vectoriels, du GMM et de la coopération des 2 modèles.

MARV	GMM	MARV & GMM
78,1 %	61,7 %	79,6 %

Ce résultat rend prometteur la mise en œuvre de systèmes hybrides basés sur la coopération de modélisations de différentes natures.

6. CONCLUSION

La reconnaissance du locuteur étant difficile à travers le réseau téléphonique, nous avons décrit une méthode permettant l'illustration des capacités d'un système de reconnaissance du locuteur. Des expériences en identification du locuteur en utilisant seulement les voyelles, diphtongues, nasales et liquides/glissées ou en utilisant les voisées donnent des résultats supérieurs à une identification utilisant les phrases entières et nous ont permis d'obtenir des résultats inégalés à ce jour sur NTIMIT: 80,2 % sur 168 locuteurs et 65,0 % sur 630 locuteurs. La compétition de modèles AR-vectoriels au niveau analytique (81,4 % sur 168 locuteurs) et la coopération des MARV et GMM sont deux approches très prometteuses. Actuellement nous étudions deux systèmes: l'un en vérification du locuteur avec phrase aléatoirement choisie imposée et contrôle du texte prononcé, l'autre basé sur la coopération des modèles AR-vectoriels et des GMM au niveau analytique.

REFERENCES

Dudani S. A. (1976) "The distance-weighted k-nearest-neighbour rule". IEEE Trans. Syst. Man Cybern., vol. SMC-6, 325-327.

Fisher W., Zue V., Bernstein J., Pallet D. (1986) "An Acoustic-Phonetic Data Base". J. Acoust. Soc. Amer. Suppl. (A), 81, S92.

Montacié C., Le Floch J.-L. (1992) "AR-Vector Models for Free-Text Recognition". ICSLP, Banff, vol. 1, 611-614.

Le Floch J.-L., Montacié C., Caraty M.-J. (1994) "Investigations on speaker characterization from ORPHÉE system technics". IEEE-ICASSP Adelaïde, vol. S_1, 149-152.

Reynolds D. A. (1994) "Speaker Identification and Verification using Gaussian Mixture Speaker Models". Workshop on autom. speaker recog. ident. verif. Proc., Martigny, 27-30.

Bimbot F., Magrin-Chagnolleau I., Mathan L. (1995) "Second-order statistical measure for text-independent speaker identification". Speech Communication, vol. 17, n° 1-2, Août, 177-192.

Le Floch J.-L., Montacié C., Caraty M.-J. (1995) "Speaker Recognition Experiments on the NTIMIT Database". Eurospeech 95 Madrid, vol. 1, 379-382.

Reynolds D. A. (1995) "Speaker identification and verification using Gaussian mixture speaker models". Speech Communication, vol. 17, n° 1-2, Août, 91-108.

Montacié C., Le Floch J.-L., Caraty M.-J. (1996) "Procédé et dispositif d'un contrôle d'accès par la voix". European patent application.

OPTIMISATION DU PARAMETRAGE ACOUSTIQUE POUR LA VERIFICATION DU LOCUTEUR

Delphine Charlet, Denis Jouvet

France Télécom CNET LAA/TSS/RCP - Technopole Anticipa, 2 avenue Pierre Marzin, 22307 LANNION Cedex

Tél.: 96 05 38 70 - Fax 96 05 35 30 - e-mail: charlet@lannion.cnet.fr, jouvet@lannion.cnet.fr

ABSTRACT

This paper presents a new framework for the optimization of the feature set in a speaker verification text-dependant DTW-based system, in which one feature set is used for time-aligning and another one is used for computing a verification score (a measure of similarity between the two acoustic forms for the given time-alignment). The feature set used for time-aligning was chosen a priori, composed of features that perform well in both speech and speaker recognition. For scoring, we propose here a method for selecting a subset of coefficients among a set of potential coefficients. The criterion used for the selection is the minimization of the experimental Equal Error Rate (EER). Experimentally, this optimization has proven to be very efficient. The selected feature subsets achieve a 15% EER reduction with or without cohort normalization. Moreover, this optimization appears to be an efficient way to reduce the feature set.

1. INTRODUCTION

Dans la plupart des systèmes de vérification du locuteur, on utilise les coefficients cepstraux, qui sont également employés en reconnaissance de parole, alors que les invariants recherchés ne sont pas du tout les mêmes dans les deux tâches. Cependant, comme les coefficients cepstraux semblent se prêter particulièrement bien aux tâches d'alignement dans les classificateurs qui les requièrent (HMM et DTW), il semble adapté de les conserver pour effectuer cet alignement. Mais pour mesurer ensuite une similarité entre des formes acoustiques mises en correspondance, en vue d'authentifier le locuteur, on souhaite utiliser un paramètrage acoustique le mieux adapté possible à la tâche de vérification. C'est pourquoi, nous proposons un nouveau système de vérification basé sur le principe de séparation des jeux de coefficients acoustiques servant à l'alignement et au calcul du score, ainsi qu'un cadre de recherche pour déterminer un jeu de paramètres optimal pour le score.

Nous présentons d'abord le système de vérification proposé, ainsi que le principe d'optimisation du paramètrage acoustique. Nous exposons ensuite différentes procédures de sélection des paramètres. Nous les comparons expérimentalement dans le dernier paragraphe où nous étudions également l'application de ce principe d'optimisation au cas du score normalisé par une cohorte d'imposteurs.

2. PRINCIPE DU CLASSIFICATEUR

Nous proposons un système de comparaison de formes acoustiques dans lequel nous dissocions les jeux de paramètres acoustiques servant aux tâches d'alignement et de calcul du score d'authentification.

Pour effectuer l'alignement qui détermine la meilleure correspondance entre deux formes acoustiques, on utilise les coefficients qui donnent de bons résultats en reconnaissance de parole: l'énergie de la trame, les 8 premiers coefficients cepstraux (MFCC) et des approximations de leurs dérivées premières et secondes (ce vecteur est appelé c27 dans la suite).

Pour le calcul du score de vérification, on cherche le jeu de coefficients qui va discriminer au mieux entre les locuteurs. La recherche d'un tel jeu de coefficients constitue à elle seule un problème, car il y a de nombreux coefficients potentiels, que l'on peut combiner d'une multitude de façons, selon de multiples critères.

Nous proposons ici un cadre de recherche pour la détermination du meilleur jeu de coefficients pour le calcul du score d'authentification. En effet, il est possible d'étudier divers coefficients potentiels, et de leur associer une mesure de " qualité " pour ensuite ne retenir que les meilleurs selon cette mesure (Bocchieri, 1992). Nous pensons qu'il est préférable de raisonner en termes de jeux de coefficients plutôt qu'en termes de coefficients étudiés séparément. Il s'agit donc de sélectionner le " meilleur " jeu parmi un ensemble de coefficients possibles. Il faut alors définir le critère qui détermine en quoi un jeu de coefficients est " meilleur " qu'un autre. Le critère que nous utilisons est le suivant: un jeu A de coefficients est meilleur qu'un jeu B s'il conduit à moins d'erreurs de classification (accepter un imposteur ou rejeter un client) que le jeu B. Il s'agit donc de rechercher le jeu de coefficients qui minimise le taux d'erreur. Dans le cas de la vérification, puisque les taux d'erreur varient en fonction du seuil de décision, nous nous intéressons à un taux d'erreur particulier, le Taux d'Egale Erreur, EER (de l'anglais *Equal Error Rate*), qui est indicatif des capacités du système. Ce critère de sélection semble assez naturel (Sambur, 1975); néanmoins, il n'avait pas été utilisé jusqu'à présent, en raison du coût de calcul prohibitif qu'il pouvait entraîner. Dans notre cas, du fait de l'implémentation particulière de notre système, nous pouvons déterminer l'EER correspondant à un jeu de coefficients particulier, en un temps assez court.

3. PROCEDURES DE SELECTION

Si l'on dispose de N coefficients potentiels pour le calcul du score, il existe 2^N sous-ensembles de coefficients possibles (si on ne fixe aucune contrainte a priori sur la taille du jeu de coefficients). La détermination du meilleur jeu de coefficients implique donc l'évaluation des 2^N sous-ensembles. Ceci est prohibitif pour des valeurs de N assez grandes (supérieures à 10). Il s'agit alors d'étudier des heuristiques de sélection de sous-ensembles de coefficients, en général sous-optimales, mais qui sous certaines conditions donnent le meilleur jeu de coefficients issu de l'étude exhaustive. Nous en présentons ici 4, qui nous ont paru intéressantes, du point de vue de leur coût en calcul et des conditions sous lesquelles ces procédures trouvent le même jeu que l'étude exhaustive.

- Procédure des N-meilleurs : on détermine un classement des coefficients selon le critère utilisé, puis on crée le jeu de n ($1 \leq n \leq N$) coefficients en considérant les n meilleurs coefficients selon le classement. Cette procédure requiert 2N évaluations du critère et suppose que le meilleur jeu de n coefficients est construit avec les n meilleurs coefficients étudiés séparément.

- Procédure de sélection ascendante : dans cette procédure, on considère que le meilleur jeu de n+1 coefficients est construit à partir du meilleur jeu de n coefficients, auquel on rajoute le coefficient parmi les N-n restants qui conduit au meilleur jeu de n+1 coefficients. Ceci nécessite N(N+1)/2 évaluations du critère.

- Procédure d'élimination progressive : il s'agit d'une procédure tout à fait symétrique de la précédente. Dans ce cas, on considère tous les coefficients, puis on retire un par un les coefficients. Cette procédure nécessite également N(N+1)/2 évaluations du critère.

- Sélection par programmation dynamique : dans cette procédure, on ne fait plus l'hypothèse que le meilleur jeu de n-1 coefficients est inclus dans le meilleur jeu de n coefficients. On détermine le meilleur jeu de n coefficients par programmation dynamique selon la méthode exposée dans (Cheung, 1978). Cette procédure nécessite $N^2(N-1)/2$ évaluations du critère.

4. EXPERIENCES

4.1. Protocole expérimental

Notre système fonctionne en mode dépendant du texte sur des phrases-clé. Il est basé sur une technique de DTW, où l'on compare l'occurrence de test à une référence du locuteur, référence construite en moyennant trois enregistrements provenant d'un seul appel. Le score résultant (dans lequel la contribution de chaque coefficient est normalisée par sa variance intra-locuteur) est comparé à un seuil prédéfini, ce qui entraine l'acceptation ou le rejet. Ce système, relativement simple de conception, est intéressant pour ses utilisateurs, du fait de la courte phase d'apprentissage requise. Son

originalité tient dans le fait qu'il utilise les coefficients acoustiques c27 pour l'alignement et un autre jeu (en l'occurrence, ici, un sous-ensemble des c27) pour le calcul du score de vérification.

Nous avons utilisé une base de données de 73 locuteurs-clients (hommes et femmes), ayant chacun enregistré deux sessions, à travers le téléphone, en général dans des conditions différentes (bureau, domicile, cabine publique, bi-bop...). Dans chacune de ces sessions, les locuteurs-clients ont répété cinq fois une série de cinq phrases-clé courtes (d'une durée inférieure à 3s chacune). Trois enregistrements provenant de l'une des sessions sont utilisés pour l'apprentissage, tandis que l'autre session est utilisée pour le test, et inversement. Nous disposons d'une base distincte de 131 imposteurs, qui ont enregistré une seule session, dans laquelle ils ont énoncé une fois les cinq phrases-clé. Par phrase-clé, nous disposons en moyenne de 2000 tests de clients et 2000 tests d'imposteurs.

Comme il s'agit ici d'évaluer les différents jeux de coefficients possibles, nous ne fixons pas de seuil a priori, nous le déterminons a posteriori, commun à tous les locuteurs, de façon à obtenir le Taux d'Egale Erreur, EER.

4.2. Résultats

4.2.1. Comparaison des procédures de sélection

Nous nous intéressons ici à la comparaison des différents algorithmes de détermination des sous-ensembles de coefficients. Ainsi, chaque procédure nous donne, pour un nombre de coefficients fixé, le " meilleur " sous-ensemble qu'elle a pu trouver. Dans la figure 1, on présente l'EER correspondant à chacun des sous-ensembles obtenus par les différentes procédures.

On observe que sur les 4 procédures présentées, les 3 dernières ont un comportement très similaire, tandis que la première (la procédure des N-meilleurs) conduit à des résultats nettement moins bons. Ainsi, construire un sous-ensemble de n coefficients en prenant les n-meilleurs coefficients considérés séparément ne donne pas, loin s'en faut, le meilleur sous-ensemble de taille n, du moins quand le critère utilisé est la minimisation du taux d'erreur.

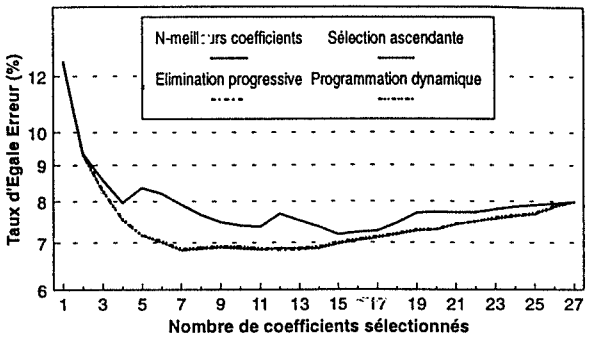


Figure 1: Taux d'Egale Erreur pour les sous-ensembles sélectionnés, selon les procédures de sélection

En ce qui concerne les 3 autres procédures, elles s'accordent sur la sélection des sous-ensembles de petite taille (lorsque l'ajout d'un coefficient diminue fortement le taux d'erreur), et sur les sous-ensembles de grande taille (lorsqu'on a déjà sélectionné tous les coefficients intéressants et qu'il ne reste plus que les coefficients dont l'ajout va dégrader les performances). On remarque une zone intermédiaire, dans laquelle l'ajout de coefficients ne modifie pas de façon significative l'EER. C'est dans cette zone que les sous-ensembles sélectionnés diffèrent selon les procédures.

Les procédures de sélection ascendante et d'élimination progressive, qui sont de complexité équivalente, ne diffèrent que dans la zone où l'ajout de coefficients n'est pas significative. La procédure par programmation dynamique, qui n'implique pas, contrairement aux autres, que le meilleur jeu de n-1 coefficients soit inclus dans le meilleur jeu de n coefficients, a pourtant, dans la majorité des cas, ce comportement. Enfin, si on s'intéresse au jeu donnant l'EER minimal, les trois procédures trouvent le même jeu de 12 coefficients (composé des coefficients cepstraux d'indices assez élevés, des dérivées premières des coefficients cepstraux, et de l'énergie). Ce jeu nous donne un EER de 6.8%, à comparer au 8.0% obtenu en utilisant l'ensemble des c27 pour le calcul du score.

4.2.2. Application au score normalisé par une cohorte d'imposteurs

Depuis (Higgins, 1991), la normalisation du score par une cohorte d'imposteurs est une technique qui s'est largement répandue. L'idée est de ne plus considérer seulement le score du test par rapport au modèle du locuteur client,

mais aussi le score du test par rapport aux modèles d'autres locuteurs. Le score normalisé est le rapport entre ces deux scores. De multiples variations autour de cette idée sont possibles (dans la construction du modèle concurrent). Le succès de cette méthode vient du fait qu'on obtient un score robuste par rapport aux variations intra-locuteur.

Nous avons inclu dans notre système de vérification un étage de normalisation du score. La cohorte d'imposteurs que nous avons utilisée est composée de 14 imposteurs (différents des imposteurs de test) choisis de façon aléatoire.

Si on aligne et calcule un score de vérification avec les c27, le score normalisé conduit à une réduction de 33% du EER par rapport à celui obtenu par le score simple (on obtient un EER de 5.4% au lieu de 8.0%).

Lorsque nous avons élaboré cette technique d'optimisation du paramètrage acoustique, nous avons travaillé avec des scores " simples " (i.e. sans normalisation). Nous avons ensuite évalué les jeux de coefficients sélectionnés (par la procédure de sélection ascendante minimisant l'EER du score simple) avec des scores normalisés (cf. Figure 2). Il ressort que le score normalisé est plus robuste à l'influence de certains coefficients qui détériorent le score simple. En effet, si on applique les mêmes jeux de coefficients (sélectionnés pour minimiser l'EER du score simple), l'EER résultant du score simple se dégrade significativement à partir d'un certain moment, tandis que l'EER résultant du score normalisé reste stable. Ainsi, il apparaît que la normalisation du score permet de résister aux " mauvais " coefficients, qui sont mis en évidence dans la minimisation de l'EER du score simple. La procédure de sélection des coefficients permet alors de réduire le nombre de paramètres à utiliser.

Finalement, en minimisant directement l'EER des scores normalisés (cf. Figure 2), on obtient une réduction du EER de 15% environ (on passe de 5.4% à 4.6%), du même ordre que la réduction du EER pour les scores simples (le jeu de coefficients trouvé contient 16 coefficients, dont 10 sont communs avec le jeu pour le score simple, la majorité des coefficients supplémentaires étant des dérivées secondes des premiers coefficients cepstraux).

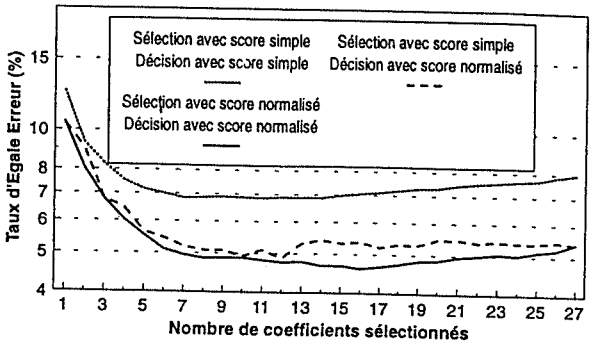


Figure 2: Taux d'Egale Erreur pour les sous-ensembles sélectionnés

5. CONCLUSION

Nous avons proposé ici un cadre de recherche pour l'optimisation du paramètrage acoustique dans un système de vérification du locuteur dans lequel on distingue les paramètres d'alignement et de score. Après avoir choisi comme critère d'optimisation la minimisation du Taux d'Egale Erreur, nous avons comparé différentes approches pour la sélection du meilleur sous-ensemble de paramètres.

Ce principe d'optimisation du paramètrage acoustique nous permet d'extraire de façon automatique les coefficients les plus intéressants d'un ensemble de coefficients potentiels, ce qui diminue significativement le Taux d'Egale Erreur et réduit efficacement le nombre de paramètres acoustiques à considérer.

6. BIBLIOGRAPHIE

Bocchieri E.L. et Wilpon J.G. (1992), Discriminative Analysis for Feature Reduction in Automatic Speech Recognition, *IEEE International Conference on Acoustics, Speech and Signal Processing*, vol.I, pp. 501-504.

Sambur M.R. (1975), Selection of Acoustic Features for Speaker Identification, *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 23, n°2, pp. 176-182.

Cheung R.S. et Eisenstein B.A. (1978), Feature Selection via Dynamic Programming for Text-Independent Speaker Verification, *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 26, n°5, pp. 397-403.

Higgins H.L., Bahler L.G. et Porter J.E. (1991), Speaker Verification using Randomized Phrase Prompting, *Digital Signal Processing*, vol. 1, n°2, pp. 89-106.

RECONNAISSANCE DE VOIX FAMILIALES

E. Perrin, J. Lescot & C. Berger-Vachon

Laboratoire Perception & Mécanismes Auditifs, UPRESA CNRS 5020, Hôpital Edouard HERRIOT, 69437 LYON Cedex 03 Tel.: 72 11 05 03 - Fax 72 11 05 34

e-mail : eperrin@cimacpcu.univ-lyon1.fr

Résumé The aim of this study was to assess a speaker recognition technic according to familial similarity of voices and to compare the performances of this technic to the judgement of three voice professionals. The automatic speaker recognition technic was based on dynamic text-dependant methods.

13 pairs of subjects, belonging to the same gender and presenting familial links were tested on two utterances of a sentence. Results are globally equivalent with acoustical and with percptual methods. Intraspeaker distances are always lower than interspeaker distance mostly when no familial links exists between the speakers.

The comparison of inter-utterance distances between acoustical and perceptual technics suggests that (1) acoustical technics discriminate between speakers, and (2) the perceptual technics sort out familial groups.

Mots-clés : Speaker recognition, Familial voices, Dynamic Methods, Perceptual judgement.

1. INTRODUCTION

La reconnissance du locuteur est un procédé de reconnaissance automatique de la personne qui parle sur la base d'informations individuelles contenue dans le signal vocal [1]. Ces systèmes de reconnaissance du locuteur sont basés sur le fait qu'entre deux locuteurs il existe des variations d'anatomie et d'habitudes vocales.

Ces variations sont moindres dans le cas de voix familiales. La comparaison ''perceptuelle'' de voix entre parents et enfants du même sexe fait souvent ressortir des phénomènes de ressemblance qui ont une base dans la transmission de caractères morphologiques de parents à enfants et d'effets de mimétisme dans l'apprentissage du langage.

La robustesse d'un système de reconnaissance automatique du locuteur devant les voix d'imposteurs doit être analysée avec soin avant toute utilisation du système autant dans les domaines commerciaux que dans les domaines de police criminelle et d'applications militaires [1]. La présentation d'imposteurs familiaux comme des jumeaux, ou tout simplement des membres d'une même famille, à des systèmes de reconnaissance automatique du locuteur a été, à notre connaissance peu étudiée. Cohen & Vaich ([2], 1994) ont étudié la présentation des voix de 10 paires de jumeaux homozygotes à un classifieur par minimum de distance. Dans au mieux 95% des cas, le système séparait les deux jumeaux alors qu'une préetude par jury d'écoute sur 5 cas avait fait ressortir 100 % de succès. Les méthodes acoustiques peuvent être abusées par un tel cas de figure.

L'étude d'impostures par voix familiales sur de tels système présente des intérêts dans les applications de reconnaissance du locuteur. Une analyse des caractères acoustiques et perceptuels correspondant à une ressemblance entre deux voix n'a pas été finement étudiée à notre connaissance. Une telle analyse permettrait d'une part de mieux comprendre les phénomènes de ressemblance vocale, autant acoustiquement que perceptuellement, et d'autre part de les appliquer à la fiabilisation des systèmes de reconnaissance automatique du locuteur.

Le but de notre étude était de confronter un système de reconnaissance automatique du

locuteur à des séparations de voix de proches parents de même sexe avec des relations de type filiation ou fratrie. Un système de reconnaissance du locuteur par analyse dynamique a été mis au point et testé sur des couples de sujets proches parents. Parallèlement, un examen perceptuel des voix a été effectué par un jury de trois professionnels de la voix sur 12 critères. Ce travail constitue la préétude d'une recherche des caractères acoustiques des voix familiales et ses applications dans la détection d'impostures vocales de type familiales et dans la recherche des caractères qui peuvent rassembler des groupes de voix, autant sur les plans objectifs que perceptuels.

2. MATERIEL ET METHODES

2.1 Bases de données

Les voix de 13 paires de sujets présentant des parentés proches de type filiation ou fratrie ont été considérés. Au sein d'une même paire, les sujets sont de même sexe. Deux prononciations d'une même phrase ont été enregistrées pour chaque locuteur : *''Alors la bise s'est mise à souffler de toute sa force mais plus elle soufflait, plus le voyageur serrait son manteau autour de lui et, à la fin, la bise a renoncé à le lui faire ôter''*. Un magnétophone REVOX B77 et un microphone REVOX 600 Ohms ont été utilisés pour les enregistrements.

Pour chaque sujet, les deux prononciations ont été distinguées en prononciation de référence (PF) et prononciation de présentation (PP). Cette distinction est aléatoire vis-à-vis de l'ordre d'enregistrement.

Les phrases ont été numérisées à 10 kHz de fréquence d'échantillonnage avec une quantification sur 16 bits. Un filtre anti-repliement de fréquence de coupure 4 kHz a été utilisé. Ces acquisitions ont été effectuées avec une station UNICE (LIMSI-CNRS). Chaque phrase a été stockée dans un fichier.

2.2 Système de reconnaissance automatique

2.2.1 Principe

Le système de reconnaissance automatique élaboré est basé sur une méthode de comparaison dynamique (Dynamic Time Warping) décrite par Furui ([3, 4], 1981). Ce système est dépendant du texte.

Soit $A = \{a_1, .., a_I\}$ l'image acoustique de longueur I de la PP du locuteur à identifier, les ai étant des vecteurs de n paramètres cepstraux d'ordre n : $a_i=\{a_{i,1}, .., a_{i,n}\}$. Si $B = \{b_1, .., b_K\}$ l'image acoustique de longueur K de la PR du locuteur de référence. D(i,k) est la distance entre les vecteurs a_i et b_k. La distance totale entre les deux locuteurs D(A, B) est donnée par l'algorithme de comparaison dynamique (DTW) qui aligne temporellement les deux représentations A et B des prononciations.

Le choix de la distance D(i,k) permet d'explorer différentes solutions.

2.2.2 Mise en oeuvre

Les paramètres cepstraux utilisés étaient basés sur une représentation LPC (Linear Predictive Coding) du spectre : les vecteurs d'analyse étaient formés de 12 LPCC (Linear Prediction Cepstral Coefficients) à l'ordre 12. Chaque vecteur cepstral a été calculé à partir d'une trame de 256 points (25.6 ms) avec un décalage de 128 points (12.8 ms) entre deux trames consécutives. Chaque trame a été préaccentuée par :

$$s(n) = s(n)-0.90*s(n-1) \qquad \text{(eq. 1)}$$

ce qui correspond à une préemphase de 90%, puis pondérée par une fenêtre de Hamming.

Chaque vecteur d'analyse a été pondéré par :

$$w(i) = 1 + 0.5\ N \sin(\pi\ i / N) \qquad \text{(eq. 2)}$$

où i est l'ordre du coefficient LPCC et N le nombre de coefficients retenus. Dans cette étude, le nombre de coefficients considéré est égal à l'ordre de la prédiction. Cette pondération a pour but d'éliminer les contributions de bas ordre, trop sensibles au

bruit, et d'ordre élevé, pas assez discriminantes [5].

Deux types de distances D(i,k) ont été calculées : dans l'espace de dimension 12 des coefficients LPCC, ces distances correspondent au normes Lp d'ordre p=1 (distance manhattan) et d'ordre p=2 (distance euclidienne) :

$$d^p_{i,k} = \sqrt[p]{\sum_{j=1}^{12} \left| a^i_j - b^k_j \right|^p} \quad \text{(eq. 3)}$$

où a^i_j et b^k_j sont les jèmes LPCC des prononciations A et B respectivement aux instants i et k. La distance globale D(A,B) entre les deux prononciations correspond à un chemin minimum calculé grâce à l'algorihme DTW.

2.3 Analyse perceptuelle

2.3.1 Critères de jugement

12 critères sur les voix ont été retenus :

a) hauteur moyenne,
b) variations de hauteur,
c) intensité,
d) débit,
e) régularité du débit,
f) nombre de pauses,
g) durée des pauses,
h) articulation,
i) voix agréable,
j) petite ou grosse voix,
k) voix douce ou agressive,
l) voix molle ou dynamique,
m) voix monotone ou mélodieuse.

Une quotation de 0 (indice minimum) à 5 (indice maximum) a été demandée pour chaque caractère.

2.3.2 Déroulement

Les 26 prononciations ont été restituées, dans un ordre aléatoire, à partir des enregistrements sur bandes magnétiques. Trois jurés (2 orthophonistes et un étudiant en thèse) ont rempli les grilles de notations. L'anonymat des sujets a été respecté par une numérotation aléatoire des prononciations.

2.3.3 Calcul des distances interprononciations

Les distances interprononciations à partir des fiches d'analyse perceptuelles ont été calculées par une métrique Lp dans l'espace à 12 dimensions des caractères perceptuels :

$$d^p_{a,b} = \sqrt[p]{\sum_{i=1}^{12} \left| c^a_i - c^b_i \right|^p} \quad \text{(eq. 4)}$$

où c^a_i est le ième caractère perceptuel du sujet a. $d^p_{a,b}$ représente la distance de la prononciation de référence de l'individu A à la prononciation de présentation de l'individu B. Les métriques d'ordre p=1 (distance Manhattan) et d'ordre p=2 (distance Euclidienne) ont été considérées.

2.4 Modes de présentation

Chacune des 13 prononciations de référence a été comparée à chacune des prononciations de présentation. Au total 169 comparaisons sont à considérer : 13 de type intralocuteur, 13 de type familial et 143 de type étranger (comparaison de la PF d'un sujet à la PP d'un sujet autre que lui-même et n'appartenant pas au même groupe familial). Seules les comparaisons entre individus de même sexe ont été considérées, ce qui diminue le nombre de comparaison de type étranger à 85.

2.5 Analyse statistique

La comparaison des distances moyennes intralocuteur (I), intrafamiliale (F) et entre étrangers (E) a été effectuée par un test t de Student de comparaison de moyennes. Les comparaisons I par rapport à F, I par rapport à E et E par rapport à F ont été effectuées pour chacune des méthodes et des distances. Au total, 12 tests ont été effectués, et le niveau de significativité a été fixé à $0.05/\sqrt{12} \approx 0.015$ pour tenir compte de cette multiplicité.

3. RESULTATS

Les figures 1 à 4 montrent les résultats obtenus pour le système de reconnaissance automatique (Figure 1 et 3) et pour l'analyse perceptuelle (Figure 2 et 4). Les deux types de distances utilisées sont présentées : distance Manhattan (Figure 1 et 2) et distance Euclidienne (Figure 3 et 4). Le tableau 1 montre la significativité des comparaisons des moyennes des types de reconnaissance en fonction de la méthode et de la distance utilisée.

3.1 Analyse globale

Une analyse globale des résultats a montré que, pour une même prononciation de référence, les distances aux prononciations de présentation étaient toujours minimales pour une comparaison intralocuteur. Cela signifie que, globalement, le système semble assez robuste vis-à-vis de présentaion de voix de proches parents.

3.2 Analyse de l'influence des voix familiales sur la reconnaissance du locuteur

Les résultats obtenus montrent qu'en moyenne les distances interlocuteur de type familial sont situées entre les distances intralocuteurs et les distances interlocuteur de type étranger. Une analyse de la significativité montre :

a) pour la méthode acoustique, la séparation des distances intrafamiliales et de type étranger est faible (figure 1 et 3), pour les deux distances, et qu'il existe seulement une différence faiblement significative avec la distance manhattan (Tableau 1),

b) pour la méthode perceptuelle, la séparation des distances intralocuteur et familiales est faible (figure 2 & 4), et il existe seulement une différence faiblement significative pour la distance euclidienne (Tableau 1).

Les autres types de comparaison (Tableau 1) font apparaître des différences hautement significatives.

3.2.1 Influence de la distance utilisée.

Au vu des résultats, il ne semble pas que l'une ou l'autre des distances présente un réel avantage dans la reconnaissance (Tableau 1). Même si la distance Manhattan semble améliorer la méthode acoustique et même si la distance Euclidienne semble améliorer la méthode perceptuelle, les niveaux de significativité sont trop faibles par rapport aux nombre de voix testées pour être pris en compte.

4. DISCUSSION

4.1 Imposture de voix familiales

Les résultats montrent qu'un système de reconnaissance automatique du locuteur basé sur des méthodes acoustiques est faiblement affecté par la présentation d'imposteurs familiaux : dans tous les cas, les distances intralocuteurs ont été trouvées inférieures aux distances intrafamiliales.

4.2 Comparaison des méthodes acoustiques et perceptuelles

Au contraire de la méthode acoustique, la méthode perceptuelle semble être piégée par les voix d'une même famille. En effet, les comparaisons I par rapport à F sont autour du seuil de significativité alors que les comparaisons avec des voix étrangères font apparaître une significativité importante ($p < 0.001$, Tableau 1). La méthode perceptuelle, basée sur les caractères présentés en 2.3.1. semble être plus efficace dans la caractérisation de groupes familiaux que dans la reconnaissance de locuteur, ce qui présente une différence notable par rapport aux méthodes acoustique.

Globalement, il apparaît des intervalles dans les distances interlocuteurs :

a) pour les méthodes acoustiques, ces intervalles apparaissent entre les comparaisons de type intralocuteurs et familiales (Figures 1 et 3), signe que les méthodes acoustiques jouent un rôle de reconnaissance du locuteur,

b) pour les méthodes perceptuelles, ces intervalles apparaissent entre les comparaisons de type intrafamiliales et étrangers, signe que ces méthodes sont bien aptes à caractériser les groupes familiaux et donc les ressemblances vocales ; les

caractères définis sur les voix semblent permettre la caractérisation de voix familiales, et une étude détaillée permettra de définir les plus pertinents dans ce sens, notamment en analysant les résultats locuteur par locuteur.

5. CONCLUSION

Un système de reconnaissance du locuteur sur méthodes acoustiques dynamiques et dépendantes du texte et une étude parallèle par jury d'écoute ont été élaborés et testés sur 13 paires de sujets proches parents.

L'analyse des résultats a montré que (1) le système acoustique semble être peu influencé par la présentation de voix familiales alors que (2) l'analyse perceptuelle a montré une plus faible robustesse dans ce domaine. Il semble que les paramètres perceptuels choisis permettent plus une reconnaissance de groupe familiaux que du locuteur.

La suite du travail permettra, sur un nombre plus important de sujets, de montrer la robustesse du système acoustique par la recherche de seuils de détection, et de déterminer les paramètres perceptuels les plus aptes à caractériser un groupe de voix familiales.

6. REMERCIEMENTS

Les auteurs remercient Marylène ARBAND et Solveig MAGNON pour leur contribution à ce travail.

7. BIBLIOGRAPHIE

Furui S. (1994) An overview of speaker recognition technology, *ESCA Workshop on Automatic Speaker Recognition, Identification and Verification, Martigny, Avril 1994*, 1-9.

Cohen A. & Vaich T.(1994) On the identification of twins by their voices, *ESCA Workshop on Automatic Speaker Recognition, Identification and Verification, Martigny, Avril 1994*, 213-216.

Furui S. (1981) Cepstral Analysis Technique for Automatic Speaker Verification, *IEEE Trans. Acoust. Speech Processing*, n°29, 254-272.

Furui S. (1981) Comparison of speaker recognition methods using statistical features and dynamic features, *IEEE Trans. Acoust. Speech and Signal Processing*, n°73, 342-350.

Juang B.H., Rabiner L.R. & Wilpon J.G.(1987) On the use of band-pass liftering in speech recognition, *IEEE Trans. Acoust. Speech and Signal Processing*, n°35, 947-954.

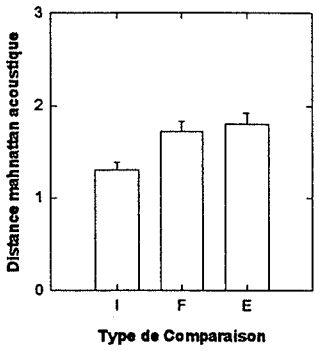


Figure 1: Comparaison des distances manhattan interlocuteur pour la méthode acoustique. I comparaison intralocuteur, F comparaison intrafamiliale et E comparaison de type étranger.

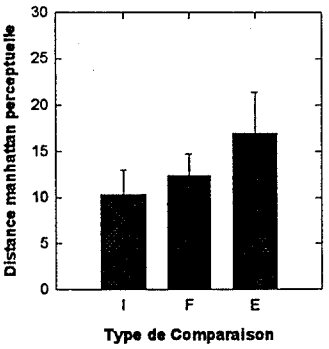


Figure 2: Comparaison des distances manhattan interlocuteur pour la méthode perceptuelle

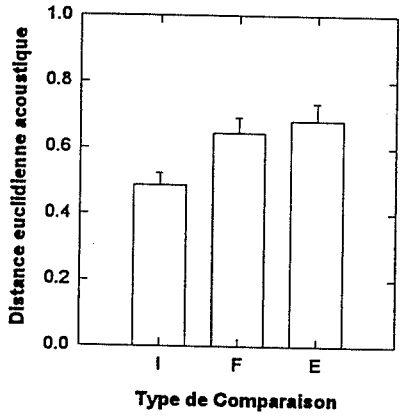


Figure 3: Comparaison des distances euclidiennes interlocuteur pour la méthode acoustique

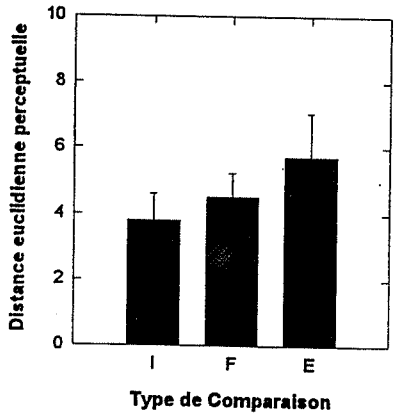


Figure 4: Comparaison des distances euclidiennes interlocuteur pour la méthode perceptuelle

Table 1: Niveaux de significativité d'un test t de Student de comparaison des distances moyennes interlocuteur : I distances intralocuteur, F distance intrafamiliales et E distance à un locuteur étranger

Comparaison	Acoustique		Perceptuelle	
Distance	Manhat.	Eucl.	Manhat.	Eucl.
I par rapport à F	**p<0.001**	**p<0.001**	p=0.023	**p=0.011**
M par rapport à E	**p<0.001**	**p<0.001**	**p<0.001**	**p<0.001**
E par rapport à F	**p=0.009**	p=0.02	**p<0.001**	**p<0.001**

STRATEGIES EN IDENTIFICATION AUTOMATIQUE DES LANGUES. VERS UNE CLASSIFICATION AUTOMATIQUE DES SYSTEMES VOCALIQUES *

François PELLEGRINO, Régine ANDRE-OBRECHT

Inst. de Recherche en Informatique de Toulouse 118, route de Narbonne - 31062 Toulouse Cedex

Tél : 61 55 60 55- Fax : 61 55 62 58 - e-mail : pellegri@irit.fr

ABSTRACT

A review of the Automatic Language Identification existing systems and of the acoustic, phonetic or prosodic features yhey use shows that they are mainly deigned to identify a language from a limited set of possibilities. In order to identify families of languages inside a more consequent set, we introduce a new approach based on vowel system identification. Our strategy consists in extracting vowel segments from the speech signal in order to identify the vowel system of the language (i.e. the number and positions of the vowels in the acoustic space) according to an existing typology.

1. INTRODUCTION

La mondialisation des communications et le caractère multi-ethnique des grandes métropoles font émerger de nouvelles applications de l'Identification Automatique des Langues.

L'intérêt d'un système rendant multilingue les serveurs vocaux ou les bornes interactives par exemple, est évident. La seule solution technologique envisageable pour un système de compréhension passe alors par la connaissance de la langue parlée par le locuteur.

Cependant, le rôle de l'IAL ne se limite pas au dialogue Homme-Machine ; elle intervient aussi comme interface possible dans le dialogue Humain-Humain. Un des exemples les plus frappants est celui du 911, le numéro d'appel d'urgence aux Etats-Unis, où les standardistes sont chargés de reconnaître la langue parlée par le correspondant et de l'aiguiller vers un interprète le comprenant et apte à traiter l'appel (1). Dans pareil cas, le délai entre l'appel et sa prise en charge peut être déterminant pour sauver des vies humaines. Disposer d'un système gérant automatiquement la redirection de l'appel vers le traducteur adéquat serait très utile.

2. LES APPROCHES EXISTANTES

La plupart des systèmes d'Identification Automatique des Langues existants sont basés sur une reconnaissance acoustico-phonétique du signal de parole. Plusieurs méthodes, issues de la reconnaissance automatique de la parole sont employées : de l'utilisation de phonèmes (2, 3, 4, 5, 6, 7, 8) à celle de macro-classes fixées *a priori* (9, 10) ou obtenues par quantification vectorielle (11a, 12, 13), la modélisation fait la plupart du temps appel à des paramètres de type cepstraux (2, 3, 4, 5, 6, 7, 8, 10, 12, 14) et à des réseaux markoviens (2, 3, 5, 6, 7, 8, 10, 12) ou neuronaux (9, 11, 14). Les systèmes décrits dans (2, 3, 7, 11a) sont développés autour d'un système de reconnaissance phonétique pour chaque langue à reconnaître : le système donnant le meilleur score correspond à la langue la plus probable. On trouvera dans (4, 11b, 15) une approche différente où un unique système phonétique commun à toutes les langues est utilisé.

De manière générale, ces systèmes ne basent pas uniquement la décision sur un critère acoustico-phonétique, et d'autres sources d'information reconnues comme pertinentes sont utilisées comme des indices de durée (4, 6) ou de nature prosodique (4, 10) respectivement basés sur une segmentation du signal et sur une étude de Fo sont intégrés.

Il apparaît que l'ajout d'un probabiliste de type N-grammes (N=2 ou 3) améliore les performances des systèmes (2, 3, 4, 6, 7, 12) même si cela implique une étude approfondie des langues à reconnaître pour obtenir des modèles fiables, étude nécessaire aussi pour les systèmes basés sur l'utilisation de modèles de mots (16). Le système présenté dans (4) fait pour sa part appel à une reconnaissance phonétique combinée à un modèle de langage et à un modèle prosodique.

D'autres approches ont été proposées par Li (14), Goodman (13) et Itahashi (17) :

Li applique des techniques issues de l'identification du locuteur en Identification des Langues. Considérant qu'il est difficile pour un modèle phonétique de prendre en

*Ce travail est réalisé dans le cadre d'un projet financé par la DRET, en collaboration avec l'ILPGA, l'ICP et le DDL de Lyon

compte les variations entre les locuteurs d'un même language, Li référence chaque locuteur du corpus d'apprentissage de chaque langue à partir des syllabes prononcées ; la langue parlée par le locuteur le plus proche de la phrase à tester (plusieurs critères de distance sont employés) est retenue.

Goodman utilise une approche encore différente puisque basée sur l'extraction des formants. A partir de F1, F2, F3 et F4 extraits sur des fenêtres de longueur constante, une quantification vectorielle est appliquée, et la distance de Mahalanobis est utilisée comme critère d'identification.

S. Itahashi et D. Liang proposent une méthode basée uniquement sur l'analyse de Fo : après avoir identifié les segments voisés du signal et avoir effectué une approximation de la courbe de Fo (par des fonctions linéaires), plusieurs paramètres statistiques sont extraits (moyenne de Fo, écart-type, corrélation avec l'energie...) et des algorithmes de classification (analyses factorielle et en composantes principales) sont appliqués.

Toutes ces méthodes (récapitulées Table 1) ont pour point commun de nécessiter systématiquement, pour chacune des langues à reconnaître, une connaissance phonologique et prosodique accompagnée d'exemples sonores (corpus d'apprentissage). Il s'agit en fait de systèmes où le nombre de candidats est restreint (typiquement une dizaine). Même si ce nombre de langues est en augmentation (25) et si les modèles de langage sont réestimés de manière automatique (15), une langue non apprise par le système ne pourra pas être identifiée.

La méthode que nous présentons au paragraphe suivant se veut plus globale. En nous plaçant dans l'espace des voyelles et en utilisant la formalisation des systèmes vocaliques, nous comptons étendre l'identification à des groupes de langues absents du corpus d'apprentissage.

3. UNE NOUVELLE APPROCHE

3.1 Présentation des systèmes vocaliques

En accord avec le principe proposé par Liljencrants et Lindblom (18) selon lequel le langage est un outil qui découle de la nécessité biologique et sociale de communiquer, différents modèles ont été proposés pour extraire des caractères universaux dans les systèmes phonologiques existants. Du modèle du contraste maximal qui impose que la distance entre les voyelles d'un système vocalique soit maximale (une dispersion maximale dans l'espace acoustique entraîne une discrimination maximale au niveau perceptif) à la Théorie de la Dispersion Focalisation proposée par (19) (le critère de dispersion est en concurrence avec un critère de focalisation visant à favoriser la présence de voyelles ayant deux formants proches), la prédictabilité des systèmes vocaliques a été étudiée.

Des études menées sur la base UPSID (20), à l'ICP (Institut de la Communication Parlée) ont abouti à l'élaboration d'une typologie des systèmes vocaliques (21) représentative des langues du monde.

Notre approche de l'IAL consiste à établir le système vocalique de la langue que l'on cherche à identifier à partir du signal acoustique. Cela nous amène à résoudre deux problèmes, à savoir la recherche automatique dans le signal des zones vocaliques et la recherche automatique de la structure du système vocalique à partir des voyelles trouvées.

3.2 Détection des noyaux vocaliques

Les méthodes appliquées au problème de la détection bruit/parole sont bien adaptées à la détection des noyaux vocaliques. Une segmentation automatique (22) permet de localiser les évènements acoustiques du signal, puis un étiquetage à partir des valeurs de la dérivée spectrale ou de l'abscisse curviligne (23) permet d'identifier les segments correspondants aux noyaux vocaliques. L'utilisation d'une segmentation *a priori* permet d'extraire la partie stable la plus étendue possible pour chaque noyau vocalique (figure 1).

La Dérivée Spectrale est calculée à l'instant *t* par la formule :

$$Ds(t) = \sum_{i=1}^{N-1} \alpha_i (E_{i+1}(t) - E_i(t))^2$$

où α_i est un coefficient de pondération des différents filtres, *N* le nombre de filtres répartis dans le domaine fréquentiel selon l'échelle Mel, et *Ei* l'énergie dans le filtre *i*.

Les coefficients α_i permettent de neutraliser les bandes de fréquences où les fortes valeurs des dérivées ne correspondent pas aux noyaux vocaliques (élimination des explosions et du Fo).

Les expériences montrent que si la dérivée spectrale est bien adaptée à la détection des voyelles orales, il est nécessaire de développer des stratégies plus complexes pour détecter et identifier les séries de voyelles secondaires (nasalisation, pharyngalisation...).

Actuellement, la détection est assurée par l'utilisation d'un seuil adaptatif calculé à partir de l'énergie moyenne du signal.

3.3 Identification du système vocalique

A partir du nuage de points correspondant aux noyaux vocaliques détectés lors de la précédente étape, la problématique est de déterminer le système vocalique correspondant.

Il est nécessaire de déterminer les voyelles présentes dans le signal à identifier puis de retrouver le système correspondant. Cette recherche s'apparente à la construction d'un dictionnaire où le nombre de classes est inconnu. Nous allons utiliser les techniques de quantification vectorielle couplées à l'algorithme de LBG-Rissanen (24), ce qui permet de générer le dictionnaire de manière automatique. Le critère *In* utilisé prend en compte la distortion globale *Dg* pour *n* classes et un facteur lié à l'entropie des observations :

$$In = \log(Dg) + np \times \frac{\log N}{N}$$

où N est le nombre d'observations et p la dimension de l'espace des observations.

Le nombre de classes retenu est le minimum de la fonction *In*.

A ces méthodes "traditionnelles" s'ajoutent des contraintes spécifiques fondées sur les connaissances a priori issues de la typologie des systèmes vocaliques : à chaque nombre de voyelles correspondent un faible nombre de systèmes, et ce nombre est connu. La prédictibilité de ces systèmes permet d'identifier celui correspondant au nuage de points (21).

Les expériences de validation du détecteur de voyelles sont actuellement en cours de réalisation.

4. PERSPECTIVES

L'originalité de notre approche repose sur la recherche d'une structure topologique dans l'espace vocalique plutôt que sur celle d'une suite d'états acoustico-phonétiques.

Notre approche met en lumière les difficultés inhérentes à la nature même des systèmes vocaliques : la détection des voyelles secondaires est complexe; de plus, la phase d'identification du système vocalique théorique doit pouvoir fonctionner à partir d'un système détecté incomplet (rien ne garantit que la totalité des voyelles soit présente lors du test). Nous envisageons d'intégrer au système la théorie de prédiction proposée par J. L. Schwartz (19).

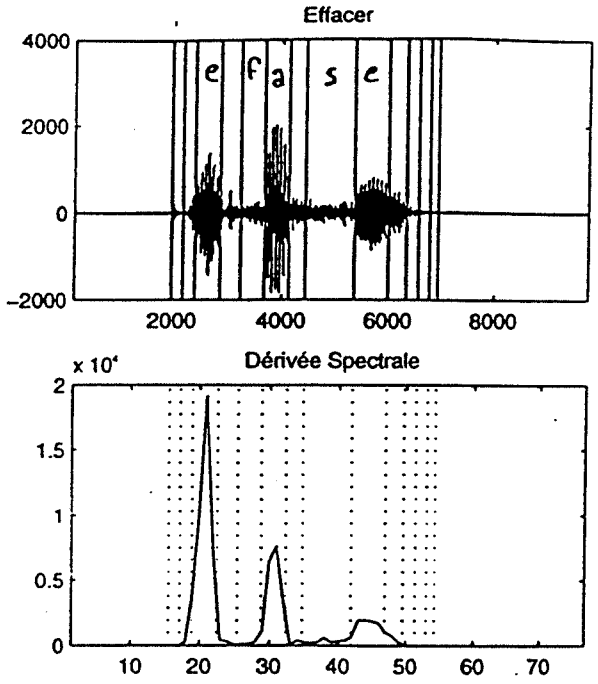


Figure 1 : Exemple de segmentation et de dérivée spectrale pour le mot "effacer"

5. BIBLIOGRAPHIE

(1) Y. K. Muthusamy, E. Barnard, R. A. Cole "*Reviewing Automatic Language Identification*" IEEE Signal Processing Magazine 10/94, pp 33-41

(2) H. Kwan, K. Hirose, "*Recognized Phoneme-Based N-Gram Modeling in Automatic Language Identification*" Eurospeech '95 Madrid, pp 1367-1370

(3)Y. Yan, E. Barnard, "*An Approach to Language Identification with Enhanced Language Model*" Eurospeech '95 Madrid, pp 1351-1354

(4) T. J. Hazen, V. W. Zue, "*Recent Improvements in an Approach to Segment-Based Automatic Language Identification*", ICSLP 94, Yokohama, pp 1883-1886

(5) P. Dalsgaard, O. Andersen, "*Application of Inter-Language Phoneme Similarities for Language Identification*" ICSLP 94, Yokohama, pp 1903-1906

(6) S. Kadambe, J. L. Hieronymus, "*Spontaneous Speech Language Identification with a Knowledge of Linguistics*" ICSLP 94, Yokohama, pp 1879-1882

(7) L. F. Lamel, J. L. Gauvain, "*Language Identification Using Phone-Based Acoustic Likelihoods*" ICASSP 94 Adelaide, pp I.293-I.296

(8) R. C. F. Tucker, M. J. Carey, E. S. Parris, "*Automatic Language Identification Using Sub-Words Models*" ICASSP 94 Adelaide, pp I.301-I.304

(9) Y. K. Muthusamy, "*Segmental Approach to Automatic Language Identification*" Ph. D. Thesis, Oregon Graduate Institute of Science & Technology, 1993

Table 1 : Récapitulatif des méthodes employées

Référence	Segment-ation	Reco. phonétique	Unités	Paramètres	Prosodie	Modèle	Corpus	Nb de langues	Résultats %	Résultats durée
Kwan 95 [2]	-	HMM	phonèmes	MFCC	-	N-grammes	OGI	5	78,00	45 s
Itahashi 95 [17]	-	-	-	Fo, durée	oui	-	OGI	6	-	-
Yan 95 [3]	-	HMM	phonèmes	MFCC	-	N-grammes	OGI	11	90,70	45 s
Lund 95 [15]	-	HMM	phonèmes	MFCC	-	N-grammes	OGI	9	-	-
Berkling 94 [11]	oui	NN	clusters	PLP	-	-	OGI	3	74,20	45 s
Hazen 94 [4]	oui	SUMMIT	phonèmes	MFCC, Fo et durée	oui	N-grammes	OGI	10	79,70	45 s
Ramesh 94 [16]	-	HMM	mots	LPC	-	-	OGI réduit	4	96,00	10 s
Dalsgaard 94 [5]	-	HMM	phonèmes	MFCC	-	-	-	4	88,10	2 mn
Kadambe 94 [6]	oui	HMM	phonèmes	MFCC, durée	-	N-grammes	OGI	3	90,00	45 s
Lamel 94 [7]	-	HMM	phonèmes	MFCC	-	N-grammes	OGI	10	59,70	10 s
Zissman 94 [12]	-	HMM	clusters	MFCC	-	N-grammes	OGI	10	79,20	45 s
Li 94 [14]	oui	NN	syllabes	MFCC	-	-	OGI	10	78,00	45s
Tucker 94 [8]	oui	HMM	phonèmes et mots	MFCC	-	-	EUROM1	3	90,00	10s
Muthusamy 93 [9]	oui	NN	classes majeures	DFT, spectre	-	-	OGI	10	62,40	45s
Savic 91 [10]	-	HMM	classes majeures	MFCC, Fo	oui	-	-	4	-	-
Goodman 89 [13]	-	-	clusters	formants	-	-	-	6	-	-

(10) M. Savic, E. Acosta, S. K. Gupta, "*An Automatic Language Identification System*" ICASSP 91 Toronto, pp 817-820

(11a) K. M. Berkling, T. Arai, E. Barnard, "*Analysis of Phoneme-Based Features for Language Identification*" ICASSP 94 Adelaide, pp I.289-I.292

(11b) K. M. Berkling, T. Arai, E. Barnard, "*Theorical Error Prediction for a Language Identification system using Optimal Phoneme Clustering*", Eurospeech 95, Madrid, pp351-354

(12) M. A. Zissman, "*Automatic Language Identification using Gaussian Mixture and Hidden Markov Models*" ICASSP 93 Minneapolis, pp II.399-II.402

(13) F. J. Goodman, A. F. Martin, R. E. Wohlford, "*Improved Automatic Language Identification in Noisy Speech*" ICASSP 89 Glasgow, pp 528-531

(14) K. P. Li, "*Automatic Language Identification using Syllabic Spectral Features*" ICASSP 94 Adelaide, pp I.297-I.300

(15) M. A. Lund, H. Gish, "*Two Novel Language Model Estimation Techniques for Statistical Language Identification*" Eurospeech '95 Madrid, pp 1363-1366

(16) P. Ramesh, D. B. Roe, "*Language Identification with Embedded Word Models*", ICSLP 94, Yokohama, pp 1887-1890

(17) S. Itahashi, L. Du, "*Language Identification Based on Speech Fundamental Frequency*", Eurospeech '95 Madrid, pp 1359-1362

(18) J. Lijencrants, B. Lindblom, "*Numerical Simulation of Vowel Quality Systems : The Role of Perceptual Contrast*", Language 48, 1972

(19) J. L. Schwartz, L. J. Boe, P. Perrier, B. Guerin, P. Escudier, "*Perceptual Contrast and Stability in Vowel Systems : A 3-D Simulation Study*", Eurospeech '89, Paris, pp 63-66

(20) I. Maddieson, "*Patterns of Sounds*", Cambridge studies in speech science and communication, Cambridge : Cambridge University Press, 1984

(21) N. Vallee, "*Systèmes vocaliques : de la typologie aux prédictions*", Thèse de Doctorat es Sciences du Langage : Modèles et Traitement en Communication Parlée, Université Stendhal, Grenoble, Octobre 94

(22) R. André-Obrecht, "*A New Statistical Approach for Automatic Speech Segmentation*", IEEE Trans. on ASSP, January 1988, vol36 no 1 pp 29-40

(23) J. B. Puel, R. André-Obrecht, "*Détection des début et fin de parole en environnement difficile*", GRETSI 93.

(24) R. André-Obrecht, "*Segmentation et Parole ?*", Habilitation à diriger des recherches, Université de Rennes IRISA, Juin 1993

(25) T. L. Lander, R. A. Cole, B. Oshika, M. Noel, "*The OGI 22 Language Telephone Speech Corpus*", Eurospeech 95, Madrid, pp817-820

JEP 96

LEXIQUE ET DIALOGUE

AVIGNON 10-14 JUIN 1996

AMELIORER LA RECONNAISSANCE DE LA PAROLE PAR L'INTEGRATION DE CONTRAINTES LINGUISTIQUES ROBUSTES : LE MODELE MICROSEMANTIQUE ALPES

Jean-Yves Antoine, Jean Caelen

Laboratoire CLIPS-IMAG — Campus Universitaire — BP 53 — F-38041 Grenoble Cedex 9
Tél: (33) 76 63 56 51 Tél: (33) 76 51 46 27
Email : Jean-Yves.Antoine@imag.fr Email : Jean.Caelen@imag.fr

1. PROBLEMATIQUE

Au cours des vingt dernières années, les recherches menées en communication orale homme-machine ont été dans une large mesure consacrées à la reconnaissance de la parole. (RdP). En effet, cette étape a longtemps été perçue comme la pierre d'achoppement du domaine. L'application de la théorie des modèles de Markov cachés (HMM) à la RdP (Rabiner 89) permet désormais d'atteindre des taux de reconnaissance très satisfaisants (Jouvet 95). D'aucuns doutent d'ailleurs de la possibilité d'améliorer significativement les performances intrinsèques des modèles actuels.

En dépit de ces résultats, il est téméraire de parler de reconnaissance robuste de la parole continue. D'une part, les systèmes de RdP sont rarement évalués dans des conditions réelles. D'autre part, les modèles markoviens de RdP présentent une combinatoire élevée qui ne peut être ignorée. Il est alors d'usage de contraindre la reconnaissance par un modèle de langage. Ce contrôle linguistique est essentiel, puisqu'il favorise l'émergence de la bonne séquence au sein d'un large treillis d'hypohthèses.

La reconnaissance de la parole continue requiert ainsi une modélisation pertinente du langage parlé. Cet article présente un analyseur linguistique (ALPES[1]) qui est précisément adapté à la parole spontanée.

2. TRAITEMENT DU LANGAGE PARLE

Dans la perspective d'une CHM orale naturelle, deux caractéristiques centrales du langage parlé sont à considérer (Blanche-Benveniste 90). Tout d'abord, ce dernier est d'une grande richesse structurelle. Ainsi, la complexité des énoncés oraux n'a rien à envier à celle de l'écrit. Ensuite, la parole spontanée présente une très grande diversité structurelle. Nombres de constructions orales (hésitations, répétitions, corrections, ellipses, incises...) échappent ainsi à toute norme grammaticale, bien qu'étant essentielles à la bonne conduite du dialogue. Leur modélisation est donc une nécessité. Malheureusement, les modèles de langage actuels ne surmontent pas la double contrainte de complexité et de diversité structurelle imposée par la parole spontanée.

Tout d'abord, la variabilité structurelle de la parole interdit toute approche syntaxique issue du TALN. C'est pourquoi l'analyse linguistique des énoncés oraux se limite généralement à une extraction d'îlots informatifs (De Mori 94). Cette approche sélective a donné des résultats très satisfaisants dans certains cadres applicatifs (réservation téléphonique par exemple). On peut néanmoins douter de sa généralisation à des domaines impliquant une communication plus riche. Ainsi, les énoncés que nous avons rencontrés dans une situation de dessin sur ordinateur (Antoine 95) ne peuvent être décrits par un simple ensemble de mots clefs.

Enfin, les modèles stochastiques de langage (De Mori 95), aisément intégrables avec une RdP markovienne, présentent malheureusement une couverture linguistique très limitée. Il ne s'adaptent donc qu'imparfaitement à la parole spontanée.

3. ALPES

Le parseur ALPES cherche précisément à modéliser la parole spontanée dans toute sa complexité et sa diversité structurelle. Il construit ainsi une structure de phrase qui reprend l'ensemble de l'information contenue dans l'énoncé, à la différence des approches sélectives. Les constructions orales inattendues sont toujours compréhensibles. C'est pourquoi nous avons basé cette analyse structurelle, qui est normalement l'apanage de la syntaxe, sur une composante sémantique lexicale qualifiée de *microsémantique* (Rastier 95). Elle repose sur un processus d'amorçage qui confère un comportement prédictif à l'analyseur. ALPES répond ainsi à deux besoins:

1. Préparer l'interprétation des énoncés en caractérisant leur structure sémantique.
2. Aider la RdP par la définition de contraintes linguistiques prédictives.

[1] Analyseur Linguistique de la Parole en Elocution Spontanée

3.1. Structure sémantique

Pour chaque énoncé, ALPES élabore un arbre de dépendance (structure sémantique) qui décrit l'ensemble des relations de sens entre les lexèmes[2] de la phrase (figure 1).

$$\begin{bmatrix} AGT = [Pred = 'on'] \\ Pred = 'dé\ placer' \\ TPS = [Pred = 'maintenant'] \\ OBJ = \begin{bmatrix} Pred = 'porte' \\ DET = [Pred = 'la'] \\ ATT = [Pred = 'rouge'] \end{bmatrix} \end{bmatrix}$$

Figure 1 — Structure sémantique :
"On déplace maintenant la porte rouge"

Ces relations prédicat-argument sont étiquetées par un ensemble de cas (Agent, Objet, Attribut, Temps...) qui correspondent à des fonctions sémantiques précises (Antoine 94, 142-157). Dans l'exemple de la figure 1, *la porte rouge* est ainsi identifiée comme l'objet sur lequel porte l'action de *dessiner*.

$$\begin{bmatrix} Pred = 'dessiner' \\ Souscat = \begin{bmatrix} AGT = /element\ /\ +\ /\ anim\acute{e}\ / \\ OBJ = /element\ / \\ LOC = /element\ /\ +\ /\ inanime\ / \\ \quad \mid\ /\ pte\ /\ +\ /\ lieu\ / \end{bmatrix} \end{bmatrix}$$

Figure 2 — cadre de sous-catégorisation de *dessiner*

L'analyse repose alors sur l'examen des propriétés relationnelles de chaque lexème: on recherche dans l'énoncé les arguments de chaque prédicat. Nous avons ainsi défini un lexique qui décrit l'ensemble des relations de sens qu'un lexème est susceptible de partager. Plus précisément, les relations que ce dernier domine[3] sont groupées dans un cadre de sous-catégorisation microsémantique. Par exemple (figure 2), *dessiner* domine une relation d'agent. Tout argument sous-catégorisé répond à une spécification sémantique précise. L'agent de *dessiner* correspond ainsi à un élément animé. La caractérisation des arguments repose sur l'analyse objective des relations de sens observées dans des corpus de dialogue naturel. En dépit de ces précautions, la généricité de la caractérisation obtenue n'est acceptable que dans le cadre restrictif d'une CHM finalisée.

La structure de l'énoncé est obtenue par unification des cadres de sous-catégorisation des prédicats et de leur arguments[4]. Cette unification obéit à quatre principes:

1. Tout argument sous-catégorisé est au plus satisfait par un lexème (ou coordination).
2. Tout lexème satisfait au plus un argument.
3. Un argument peut éventuellement être vide.
4. On ne peut coordonner que des lexèmes de même cas.

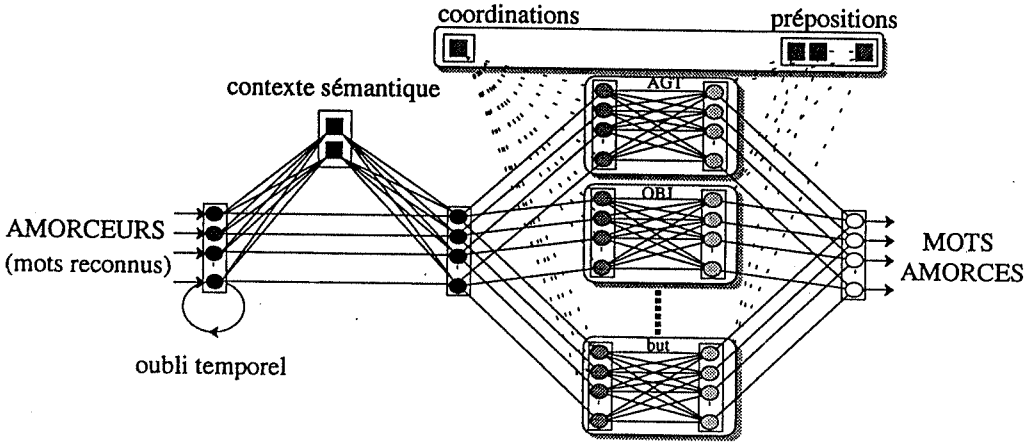


Figure 3 : Réseau associatif d'amorçage sémantique

3.2. Amorçage sémantique

A tout instant, l'amorçage sémantique prédit l'ensemble des lexèmes susceptibles de survenir dans la suite de l'énoncé. On analyse pour cela les relations de sens que peuvent partager les mots déjà reconnus. Compte-tenu des principes d'unification, ce processus tente uniquement de satisfaire les arguments déjà sous-catégorisés mais non encore satisfaits dans l'énoncé.

L'amorçage est réalisé en deux étapes. Tout d'abord, une phase d'*adaptation contexuelle* focalise l'amorçage sur le contexte sémantique courant. Ensuite, les amorcés sont recherchés à proprement parler (*amorçage relationnel*).

3.3. Réseau d'amorçage

Le processus d'amorçage est réalisé par un réseau associatif multicouches (figure 3) qui est

[2] lexème : unité lexicale de sens.
[3] i.e. pour lesquelles il joue le rôle de prédicat.
[4] Un argument peut bien sûr jouer aussi le rôle de prédicat.

compilé à partir du lexique sémantique. Chaque couche répond ainsi à une motivation bien établie, ce qui favorise le respect de contraintes linguistiques (principes d'unification...) par le réseau. Chaque noeud représente un lexème. Les entrées correspondent aux mots reconnus (*amorceurs*). Leur activité est propagée jusqu'à la couche de sortie, où les noeuds d'activations maximales forment l'ensemble des lexèmes *amorcés*. L'activité de sortie $a\omega$ d'un mot amorcé ω correspond alors à sa probabilité sémantique $P(\omega_i = \omega \mid \omega_{i-1}\ \omega_{i-2} \ldots \omega_0)$ d'apparition à la suite d'une séquence $(\omega_{i-1}\ \omega_{i-2}\ \ldots\ \omega_0)$ déjà reconnue.

Les équations qui régissent la propagation au sein du réseau sont détaillées dans (Antoine 96). Nous nous contenterons ici d'évoquer rapidement le déroulement de l'amorçage.

Tout d'abord, les entrées sont soumises à un processus d'oubli temporel qui favorise les amorceurs récents sans interdire néanmoins les amorçages à longue distance.

L'adaptation contextuelle est ensuite mise en oeuvre dans les trois premières couches du réseau. Elle se traduit par une pondération des entrées en fonction du domaine sémantique courant. Chaque lexème est associé à un champ sémantique donné, correspondant à un noeud de la seconde couche: le contexte sémantique courant est donné par le noeud de plus haute activation. La troisième couche correspond aux entrées pondérées par la couche contextuelle.

L'amorçage relationnel est ensuite réalisé en parallèle sur plusieurs sous-réseaux casuels (un réseau par cas) afin de respecter les contraintes d'unification. Nous devons en effet distinguer les différents arguments des lexèmes. Ainsi, les entrées sont réparties suivant les différents réseaux casuels (couche 4). Puis, on recherche pour chaque argument (i.e. cas sémantique) les amorcés les plus intéressants. La couche 5 correspond ainsi aux activations casuelles des amorcés. On regroupe enfin ces activations dans la couche de sortie, pour récupérer les probabilités globales d'amorçage sémantique.

Le réseau est contrôlé par le processus d'unification. Par exemple, dès qu'un argument sous-catégorisé est instancié, le noeud qui lui correspond dans la couche 4 est mis à zéro. De même, ce contrôle concerne:

1. Prépositions — l'amorçage relationnel est limité aux sous-réseaux qui correspondent à un cas compatible avec la préposition.

2. Coordinations — l'amorçage relationnel est limité aux relations de sens déjà instanciées (principe 4 appliqué à l'analyse gauche-droite). Répétitions, reprises et corrections suivent cette règle, même en l'absence de coordonnant.

4. RESULTATS

4.1. Couverture linguistique

Sans recours à la syntaxe, ALPES réalise un *parsage* complet des énoncés (extraction de l'ensemble des relations structurelles, gestion des coordinations et des prépositions, etc.). Il possède ainsi une couverture linguistique très appréciable, qui concerne en particulier les constructions suivantes (Antoine 96):

* passifs
* questions ouvertes ou fermées.
* subordonnées, relatives, complétives.
* présentatifs: *il y a un carré à gauche.*
* extractions: *le carré on le déplace.*

Puisque l'analyse microsémantique est indépendante de la forme grammaticale des énoncés, ALPES surmonte en outre la plupart des inattendus de l'oral (Antoine 96):

* hésitations
* répétitions, reprises
* auto-corrections
* interruptions

Seules posent problème certaines incises et ellipses étendues. Leur résolution ne peut en fait être envisagée qu'à un niveau pragmatique.

4.2. Robutesse

La robustesse d'ALPES a précisément été confirmée par plusieurs études menées sur des corpora[5] de parole spontanée. Ces expériences visaient à comparer ALPES avec *Ln2_3*, un analyseur à grammaire lexicale fonctionnelle développé pour l'écrit (Zweignebaum 91).

Table 1 — Robustesse moyenne des deux analyseurs (nb. d'analyses correctes / nb. total d'énoncés de test).

Corpus	Ln2_3	ALPES
Corpus 1	0,408	0,853
Corpus 2	0,401	0,785
Corpus 3	0,767	0,866
Moyenne	*0,525*	*0,835*

La Table 1 montre clairement l'intérêt de notre approche. On constate en effet que les performances d'ALPES sont très sensiblement supérieures à celles de *Ln2_3*.

Tableau 2 — Etude inter-individuelle de robustesse sur 6 locuteurs du corpus 1.

Robustesse	Ln2_3	ALPES
Minimale	0,18	0,68
Maximale	0,74	0,96
Ecart-type	*0, 20*	*0,09*

[5] Ces expériences ont été réalisées à partir de la transcription écrite littérale de 504 énoncés oraux rencontrés en situation naturelle de communication homme-homme finalisée.

En outre, ALPES présente une robustesse relativement constante sur les 3 corpora, à la différence de la LFG. On retrouve cet écart de stabilité en analysant le comportement des deux analyseurs pour différents locuteurs (Table 2).

5. RECONNAISSANCE ROBUSTE

Dans l'optique d'un contrôle linguistique sur la RdP, les performances d'ALPES sont particulièrement intéressantes. Nous étudions ainsi l'intégration de ce parseur à un processus de reconnaissance markovien (Delemar 95). Deux modes d'intégration sont envisagés.

5.1. Filtrage post-reconnaissance

Cette intégration est limitée au filtrage du treillis de reconnaissance, afin de ne conserver que les séquences sémantiquement cohérentes. Se pose alors la question du classement des séquences conservées: quel crédit accorder à un énoncé de cohérence limite mais présentant un score de reconnaissance élevé? Nous proposons de pondérer ce dernier score par un score d'intelligibilté qui correspond au produit des probabilités d'amorçage[6]:

$$P(\omega) = \left(\prod_i P\mu sem(\omega_i | \omega_{i-1}\ \omega_{i-2} \ldots\ \omega_0) \right)^{1/n}$$

Le score global $P(\omega|X)$ d'une séquence de mots ω, pour une observation X, est alors:

$$P(\omega|X) = \frac{P(X|\omega).P(\omega)}{\sum_i P(X|\omega_i).P(\omega_i)}$$

où $P(X|\omega)$ est la probabilité d'émission de X compte-tenu de la suite des HMMs lexicaux correspondant à la séquence ω.

S'il permet un rejet des séquences non cohérentes, ce filtrage tardif ne garantit pas la présence finale de la bonne solution, qui a pu être préalablement écartée par la RdP. C'est pourquoi nous envisageons une interaction plus précoce entre RdP et analyse linguistique.

5.2. Interaction précoce

L'idée est ici de contraindre dynamiquement l'enchaînement des HMMs lexicaux à l'aide des probabilités d'amorçage. A l'opposé des N-grams, ALPES a un domaine de contextualité étendu à l'ensemble de l'énoncé. Cette faculté, essentielle à une modélisation linguistique de qualité, permet ainsi de rejetter l'énoncé (b) :

(a) *Je mange souvent du très bon **pain**.*
(b) * *Je mange souvent du très bon **whisky**.*

Contrairement aux modèles stochastiques de langage, ALPES ne peut par contre définir ces contraintes sémantiques que d'une manière dynamique[7]. Bien que théoriquement viable, cette solution est très couteuse pour l'analyse linguistique (un analyse parallèle par chemin du treillis!). D'où la nécessité de filtrer à chaque instant les chemins partiels de probabilités peu élevées. Du fait de l'algorithme de Viterbi (Forney 73) , ces probabilités n'ont cependant de pertinence réelle qu'en fin d'analyse. Il nous faudra donc être très prudent dans l'élagage des séquences.

6. CONCLUSION

Nous avons présenté un parseur sémantique (ALPES) qui réalise une analyse du langage parlé à la fois détaillée et robuste. Basé sur un un processus prédictif d'amorçage sémantique, cet analyseur peut être utilisé efficacement pour contraindre une RdP markovienne. Deux modes d'intégration ont été proposés, qui visent une amélioration sensible, en terme de robustesse, de la reconnaissance de la parole continue.

BIBLIOGRAPHIE

Antoine J.Y. (1994) *Coopération syntaxe-sémantique pour la compréhension automatique de la parole spontanée*, Thèse, INPG, Grenoble.

Antoine J.Y. (1995), *Conception de dessins et CHM : améliorer l'interaction orale au niveau linguistique*, in Caelen J. et Zreik K. (ed), *Le Communicationnel pour concevoir*, Europia, Paris.

Antoine J.Y. (1996), *Parsing spontaneous speech without syntax*, COLING, Copenhague (à paraître).

Blanche-Benveniste C. et al. (1990) *Le français parlé*, CNRS Editions, Paris, France.

Delemar O., Kabre H. (1995), A bottom-up hybrid method for isolated words recognition, *ICPhS*, Stockholm, Suède, vol. 4, 268:271.

De Mori R. (1994) Apprentissage automatique pour l'interprétation sémantique, *20° Journées d'Etudes de la Parole*, Trégastel, 11:19.

De Mori R. (1995) Modèles stochastiques de langage, école d'été *Fondements et perspectives en TAP*, Marseille, 109:118.

Forney G.D. (1973) The Viterbi algorithm, *Proc. IEEE*, 61, 268:278.

Jouvet D. (1995) Modèles de Markov pour la reconnaissance de la parole, école d'été *Fondements et perspectives en TAP*, Marseille, 99:108.

Rastier F. (1995,) *La microsémantique*, in Rastier F., Cavazza M., Abeillé A, *Sémantique pour l'analyse*, Masson, Paris, 43:82.

Rabiner L.R. (1989) A tutorial on Hidden Markov Models and Selected Applications in Speech Recognition, *Proc IEEE*, 77(2), 257:285.

Zweigenbaum P. (1991) Un analyseur pour grammaires LFG,*TA Informations*, 2, 34.

[6] L'exposant en 1/n permet une normalisation du score d'intelligibilité indépendante de la longueur de la séquence.

[7] Plus précisément, une définittion statique implique une génération de l'ensemble des séquences envisageables. Cette solution est limitée à des applications très restreintes.

COMPRÉHENSION ET ÉVALUATION DANS LE DOMAINE ATIS

Wolfgang MINKER, Samir BENNACEF

LIMSI-CNRS - BP 133, 91403 Orsay cedex
email: {minker,bennacef}@limsi.fr

ABSTRACT

A spoken language system for vocal access in French to the Air Travel Information Services (ATIS) task has been developed at LIMSI. The semantic analysis in this system is based on a case grammar. We have ported the natural language understanding component to American English using the ARPA ATIS corpus. We discuss the advantages and the performance limitations of the approach when applied to an increased domain coverage and to other languages.

INTRODUCTION

Un système d'interaction verbale dans le domaine ATIS (Air Travel Information Services) permet à son utilisateur d'acquérir dans des conditions de simulation des informations issues d'un guide officiel des compagnies aériennes américaines et canadiennes [1]. Le domaine ATIS est une des applications choisies par différents laboratoires américains fédérés dans un programme lancé par l'agence ARPA (Advanced Research Project Agency) pour le développement de systèmes du traitement du langage écrit et parlé, pour la collecte de données et la mise à disposition d'outils pour l'évaluation des systèmes.

Un système d'interaction verbale pour le français, L'ATIS, a été développé au LIMSI [2]. Son architecture décrite sur la figure 1 intègre différentes composantes: reconnaissance vocale, compréhension, dialogue ainsi que génération de la requête SQL (System Query Language) et de la réponse du système. Le module de reconnaissance vocale transforme la parole en entrée en une suite de mots qui est transmise à l'analyseur sémantique. Celui-ci détermine la signification de la phrase et établit une représentation sémantique appropriée. La gestion du dialogue complète cette représentation en utilisant l'historique ou en posant des questions à l'utilisateur. Le résultat est utilisé par le module de génération de la requête pour construire une séquence de commandes SQL et extraire les informations de la base de données. La génération de la réponse s'appuie sur les résultats fournis par la base de données et sur la représentation sémantique complétée.

L'analyse sémantique dans L'ATIS est fondée sur la grammaire des cas [3]. Après une présentation de l'originalité de l'approche, nous décrirons les résultats obtenus récemment dans le domaine de la compréhension, et présentons une étude de flexibilité de la méthode pour l'augmentation de la couverture du domaine ainsi qu'une analyse de portabilité vers d'autres langues. En utilisant le corpus ARPA ATIS [4], l'analyseur sémantique de L'ATIS a été adapté à l'anglais-américain et évalué avec des données de test officielles.

REPRÉSENTATION SÉMANTIQUE

Dans le domaine de la demande d'informations, l'interaction homme-machine est spontanée, ce qui peut se manifester par des répétitions, des hésitations ou des requêtes disloquées (ruptures de construction) qui ne respectent pas la grammaire de l'écrit. L'extraction sémantique ne peut donc pas s'appuyer totalement sur l'analyse syntaxique nécessairement incomplète, mais doit se limiter aux éléments porteurs de sens de la requête tout en ignorant les parties redondantes ou non-essentielles pour l'application. La compréhension dans L'ATIS utilise cette approche fondée sur l'utilisation d'une grammaire des cas permettant de détecter dans chaque requête le ou les concepts liés à l'application et de les valider par un jeu de contraintes.

Tableau 1: Concepts utilisés par l'analyseur sémantique du système français L'ATIS. Les mots-clés soulignés dans les exemples facilitent l'identification du concept.

Concept	*Exemple*
vol	*Je voudrais aller de Oakland à Denver*
tarif	*Je voudrais les tarifs des vols de Denver à Atlanta*
escale	*Quel est le lieu de l'escale*
type	*Quel est le type d'avion pour la compagnie American*
réserver	*Je choisis le vol numéro 317*

Par exemple, dans la phrase *J'aimerais aller de Washington à Boston le cinq juin*, le concept est vol et les contraintes sont ville-départ, ville-arrivée et date. D'un point de vue de la grammaire des cas, le concept correspond à la structure casuelle alors que les contraintes correspondent aux cas.

Pour déterminer les concepts et leurs contraintes, un corpus français qui couvre un sous-domaine de l'application a été utilisé [5]. Comme le montre le tableau 1, cinq concepts ont été identifiés à partir de l'analyse manuelle de 655 phrases issues de ce corpus et associés à un ensemble de mots clés tels

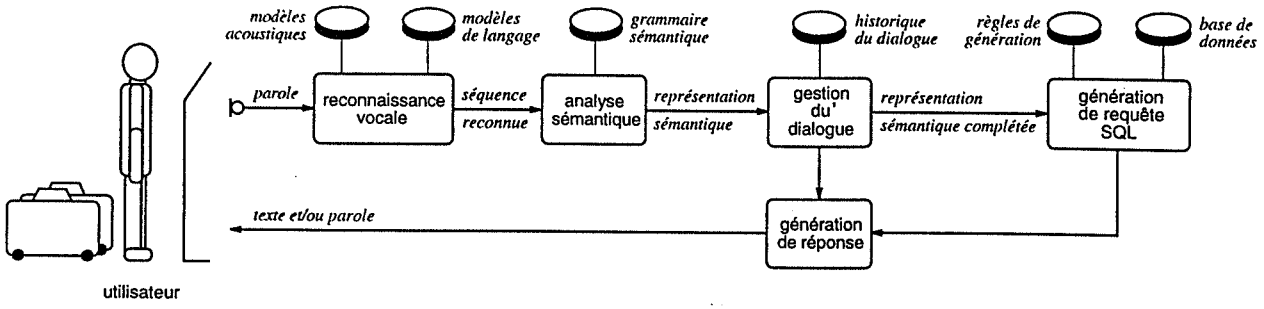


Figure 1: Architecture du système d'interaction verbale L'ATIS.

que ceux soulignés dans les exemples. Un ensemble de 38 contraintes sémantiques représente les catégories d'information, très souvent communes à tous les concepts, par exemple itinéraire, désignation-vol, durée-vol, tarif et ville. Des marqueurs de cas constituent les contraintes syntaxiques nécessaires pour l'extraction d'une représentation sémantique. Dans la phrase *de Washington à Boston* la préposition *de* désigne *Washington* comme une ville de départ et *à* désigne *Boston* comme une ville d'arrivée.

Dans L'ATIS, la grammaire est décrite dans un langage défini dans un fichier déclaratif. Elle contient l'ensemble des structures casuelles et des cas liés à l'application. L'analyse d'une phrase consiste en l'identification d'une structure casuelle pour construire une représentation sémantique sous forme de schémas. Les attributs des schémas sont instanciés à partir de certaines parties de la phrase en utilisant les marqueurs de cas.

La figure 2 montre les structures établies à différents niveaux d'analyse, la représentation sémantique (*RS*), la séquence des commandes SQL (*S*) et la réponse du système (*R*). Pour la requête *RQ*, l'analyseur choisit le concept vol, reconnu par le mot clé *aller*. La représentation sémantique *RS* se construit en instanciant les attributs de, à et escale-à avec respectivement *Philadelphie, San Francisco* et *Dallas*. La séquence *S* des commandes SQL permet d'accéder à la base de données et d'en extraire l'information. Dans l'exemple, la commande SELECT from-airport, to-airport, departure-time, arrival-time FROM flight se construit à partir du concept instancié vol de *RS*. Si les attributs de, à et escale-à contiennent des valeurs, les parties @de, @à et @escale-à dans *S* sont remplacées respectivement par les valeurs des attributs de, à et escale-à. Avant d'être présentée à l'utilisateur, l'information est mise en forme et accompagnée d'une réponse *RP* en langage naturel éventuellement synthétisée.

DONNÉES ET ÉVALUATION - UN EFFORT MULTI-SITE

En 1991, le groupe MADCOW (Multi-site ATIS Data Collection Working Group) de coordination pour la collecte multi-site d'un nombre important de données destinées au développement des systèmes en traitement du langage parlé [4] s'est créé au sein de la communauté ARPA. Le groupe a également élaboré un paradigme d'évaluation standardisée en parole et langage naturel dans le domaine ATIS [6, 1].

Dans la plupart des cas les enregistrements du corpus ont été réalisés de manière classique en utilisant les techniques du Wizard of Oz (Magicien d'Oz) où l'on fait croire aux sujets qu'ils parlent à un système complètement automatique. Des scénarios ont été proposés aux candidats sur la base desquels ils devaient formuler leurs demandes.

Pour la mise en forme, les requêtes ont été orthographiquement transcrites en utilisant des conventions de référence. Lors du processus dit d'annotation, les phrases du corpus ont été étiquetées conformément aux standards proposés par la communauté ARPA [4]. On distingue entre les requêtes de type A (indépendantes du contexte), de type D (dépendantes du contexte) et de type X (qui regroupe les phrases aberrantes). Pour chaque requête de types A et D, des réponses de référence de la base de données ont été établies. Ces références sont utilisées dans une évaluation standardisée [6]. Le corpus pour ATIS consiste en 12 000 requêtes de parole avec des transcriptions orthographiques exactes, des fichiers contenant leurs catégories, les séquences d'entrée du Magicien à NLParse[1], les séquences de commandes SQL pour les réponses de référence ainsi que les références correspondantes. Le corpus a été divisé en deux parties, développement et test. Des tests officiels ont eu lieu entre 1990 et 1994 [7].

Un objectif supplémentaire de MADCOW est la création d'une infrastructure pour l'évaluation automatique des systèmes afin de simplifier le développement et d'offrir l'opportunité de comparer des résultats entre les sites. MADCOW a proposé d'effectuer l'évaluation à partir de la réponse du système plutôt qu'à partir de la représentation sémantique. La méthode consiste à comparer une paire de réponses de référence minimale et maximale construites manuellement au préalable à une hypothèse générée par le système [6, 1].

[1]NLParse est un produit de Texas Instruments pour l'accès à une base de données

RQ	*Je veux aller de Philadelphie à San Francisco avec escale à Dallas*
RS	<vol> de: philadelphia à: san-francisco escale-à: dallas
S	SELECT airline_code, flight.flight_id, flight.departure_time, flight.arrival_time, stops, stop_airport FROM flight, flight_stop WHERE from-airport=@de AND to-airport=@à AND stop-airport=@escale-à
RP	Voici les vols de Philadelphie à San Francisco faisant escale à Dallas

COMPAGNIE	NUM_VOL	DÉPART	ARRIVÉE	ESCALES	VILLE_ESCALE
DELTA	217/149	08h30	13h25	1	DALLAS/FORT-WORTH
AMERICAN	459	15h00	20h23	1	DALLAS/FORT-WORTH
DELTA	589/395	19h15	23h50	1	DALLAS/FORT-WORTH

Figure 2: Structures établies dans le système d'interaction verbale L'ATIS. Requête (RQ), représentation sémantique (RS), séquence de commandes SQL (S) et réponse formatée (RP).

L'ATIS EN ANGLAIS-AMÉRICAIN

Adapter la partie compréhension et génération de la requête du système français L'ATIS à l'anglais-américain revient à traduire les fichiers déclaratifs contenant la grammaire des cas et les règles de la génération de requête. La version anglaise a été développée en utilisant les 3 275 requêtes du type A du corpus ATIS [4]. Lors du développement, les règles étaient modifiées, puis l'analyseur sémantique évalué de façon itérative [8]. Pour le français, la collecte de données décrite dans [5] a permis de définir un sous-domaine de l'application en fonction de l'importance accordée par la population visée à un concept sémantique particulier.

Tableau 2: Les concepts utilisés dans le domaine ATIS et quelques mots clés utilisés par l'analyseur sémantique du système anglais de L'ATIS.

Concept	*Mots clés*
abbreviation	abbreviation, define, explain
quantity	capacity, count, number of
meal	eat, food, meal
ground-service	transportation, ground
airport	airport, airports
airfare	airfare, cost, price, rate, ticket
aircraft	description, kind, type
flight-class	class
restriction	restriction
airline	airline
city	city, where
time-zone	time zone
flight	flight, operate, run, travel, trip

Le concept time-zone, par exemple, n'est pas inclus dans le système français car il n'est pas utilisé. Par contre dans le domaine ATIS, tel que défini par ARPA, les données et la couverture du domaine sont imposées. C'est pourquoi des concepts supplémentaires, meal, ground-service, time-zone ont été ajoutés. Le tableau 2 montre les 13 concepts sémantiques accompagnés de quelques mots clés déterminés pour l'anglais. Un total de 69 contraintes représente presque le double par rapport à la version française. Pour la désignation du vol, par exemple, le système français ne contient que les contraintes numéro du vol et compagnie. Dans la version anglaise, des contraintes supplémentaires existent: type d'avion et capacité. A partir de la transcription, des marqueurs et des mots-clés sont ajoutés, enlevés ou leur succession est changée dans la grammaire. Par exemple, en français, *midi* et *minuit* ne se comportent pas exactement comme des nombres. Contrairement à l'anglais, ils peuvent être directement suivis de minutes comme dans la phrase *demain midi trente*. C'est ainsi qu'ils servent seulement dans le système français comme des prémarqueurs pour l'attribut départ-minute.

PERFORMANCES ET LIMITATIONS

En utilisant une métrique standardisée introduite par la communauté ARPA [6], l'analyseur sémantique anglais de L'ATIS a été évalué sur les données de test officielles en Février 1992. En s'appuyant sur la réponse du système, la méthode consiste à comparer l'hypothèse à une paire de réponses de référence minimale et maximale. On obtient un résultat de 81,8% de réponses correctes.

Le tableau 3 montre des exemples de requêtes qui échouent systématiquement. La suite des symboles entre parenthèses permet de répertorier la requête dans le corpus. La complexité du système, par exemple, devrait augmenter considérablement pour tenir compte des concepts marginaux et pouvoir répondre à la requête *RQ1*. Le système est également incapable d'analyser correctement des requêtes contenant des contraintes multiples ou plus d'une seule catégorie sémantique (*RQ2, RQ3*). Le traitement des auto-annulations qui est important dans un système conçu pour des applications réalistes, reste le problème le plus difficile à résoudre. Bien souvent il n'est pas évident de savoir quelle partie sémantique devrait être annulée (*RQ4*). Le problème des répétitions, par contre, est maîtrisable. Dans la plupart des cas, il peut être suffisant de supprimer un des mots identiques (*RQ5*).

Dans le domaine du trafic aérien, les nombres jouent un rôle important. Ils peuvent être des dates, heures, numéros de vols, nombre d'escales ou nom-

Tableau 3: Quelques requêtes problématiques pour l'analyseur sémantique du système anglais de L'ATIS. La suite des symboles entre parenthèses permet de répertorier la requête dans le corpus.

Requête		*Type d'erreur*
RQ1	*What's the next smallest plane after a turboprop* (fb0092sx) (Quel est l'avion de taille inférieure à une turboprop)	concept marginal
RQ2	*List all the flights from Denver to Pittsburgh and list the fares* (b200a1sx) (Montrez tous les vols de Denver à Pittsburgh et montrez les tarifs)	deux concepts
RQ3	*Which airline serves Denver Pittsburgh and Atlanta* (4a00b1sx) (Quelle compagnie dessert Denver Pittsburgh et Atlanta)	contraintes identiques multiples
RQ4	*Show me flights from San Francisco from Pittsburgh to San Francisco on Monday* (j50021sx) (Montrez-moi des vols de San Francisco de Pittsburgh à San Francisco lundi)	auto-annulation
RQ5	*I need to fly from Dallas to San Francisco and be in San Francisco by four p m* (i100a5sx) (Je devrais aller de Dallas à San Francisco et être à San Francisco aux alentours de seize heures)	répétition
RQ6	*Show me flights from Baltimore to Philadelphia arriving after twenty one hundred* (rc0073sx) (Montrez-moi des vols de Baltimore à Philadelphia arrivant après vingt et une heure)	contraintes faibles pour l'heure

bre de personnes. En anglais, c'est le contexte qui peut être décisif pour attribuer un nombre à une heure (*RQ6*). Ceci entraîne un risque de confusion, car si l'on instancie un cas, des ambiguités surgissent si les marqueurs et/ou les mots clés manquent.

RÉSUMÉ ET PERSPECTIVES

Cet article décrit les résultats de l'adaptation à l'anglais-américain de l'analyseur sémantique développé initialement pour le système d'interaction verbale français L'ATIS. L'analyseur utilise une grammaire des cas afin d'extraire le contenu sémantique d'une requête formulée par l'utilisateur. Cette technique est jugée plus adaptée pour des applications dans le domaine d'interaction homme-machine spontanée, que les méthodes fondées sur une analyse purement syntaxique. L'extraction sémantique s'appuie cependant sur des éléments syntaxiques (marqueurs) ce qui rend la méthode robuste face aux effets du langage parlé ou aux requêtes dont la syntaxe n'est pas connue du système. Les connaissances sémantiques sont acquises au préalable lors de l'analyse manuelle d'un corpus de développement.

L'approche proposée est suffisamment flexible ce qui permet d'augmenter la couverture du domaine en ajoutant des concepts supplémentaires. La grammaire des cas est également portable vers d'autres langues car il suffit de traduire le jeu de règles en tenant compte de quelques spécificités de la langue. Introduite par la communauté ARPA, l'évaluation standardisée a permis de tester de façon itérative les adaptations nécessaires du système et permet donc d'en simplifier le processus. L'analyse de quelques requêtes problématiques donne une idée des limitations de l'approche. L'augmentation de la couverture du domaine améliore les performances du système, mais également sa complexité. Et dans ce cas, si l'on essaie de formaliser la grammaire, celle-ci ne peut jamais tenir compte de tous les concepts sémantiques tels que ceux imaginés par l'utilisateur. L'approche se montre également limitée face aux phénomènes qui ne sont pas accompagnés de contraintes syntaxiques parmi lesquels notamment les annulations.

Une alternative à un système fondé sur des règles peut consister en un apprentissage automatique de concepts à partir d'un grand nombre d'interactions. Un tel système fondé sur un modèle stochastique issu de l'analyse automatique d'un corpus d'apprentissage, peut intégrer les concepts et les contraintes en fonction de leur occurrence. Ce système est en cours de développement.

BIBLIOGRAPHIE

[1] P. Price. Evaluation of Spoken Language Systems: The ATIS Domain. In *Proceedings of ARPA Workshop on Human Language Technology*, pages 91–95, June 1990.

[2] S. K. Bennacef, H. Bonneau-Maynard, J. L. Gauvain, L. F. Lamel, and W. Minker. A Spoken Language System For Information Retrieval. In *Proceedings of ICSLP*, pages 1271–1274, September 1994.

[3] Ch. J. Fillmore. The case for case. *Universals in Linguistic Theory*, pages 1–90, 1968.

[4] MADCOW. Multi-Site Data Collection for a Spoken Language Corpus. In *Proceedings of DARPA Speech and Natural Language Workshop*, pages 7–14, February 1992.

[5] H. Bonneau-Maynard, J. L. Gauvain, L. F. Lamel, J. Polifroni, and S. Seneff. A French version of the MIT-ATIS System. In *Proceedings of the European Conference on Speech Technology, EUROSPEECH*, pages 2059–2062, September 1993.

[6] M. Bates, S. Boisen, and J. Makhoul. Developing an Evaluation Methodology for Spoken Language Systems. In *Proceedings of DARPA Speech and Natural Language Workshop*, pages 102–108, February 1992.

[7] D. S. Pallett, J. G. Fiscus, W. M. Fisher, J. Garofolo, B. A. Lund, A. Martin, and M. A. Przybocki. 1994 Benchmark Tests for the Spoken Language Program. In *Proceedings of ARPA Workshop on Human Language Technology*, pages 5–36, November 1994.

[8] W. Minker. An English Version of the LIMSI L'ATIS System. Technical Report 9512, LIMSI-CNRS, April 1995. Notes et Documents LIMSI.

INTÉGRATION DE DIFFÉRENTS NIVEAUX LINGUISTIQUES POUR LE TRAITEMENT DES MOTS HORS-DICTIONNAIRE DANS LA CONVERSION GRAPHÈME-PHONÈME AUTOMATIQUE

Frédéric BÉCHET, Marc EL-BÈZE

Laboratoire Informatique - 339, chemin des Meinajariès - BP 1228 - 84911 AVIGNON Cedex 9

Tél.: (33) 90 84 35 00 - Fax : (33) 90 84 35 01 - e-mail: fred,elbeze@univ-avignon.fr

ABSTRACT

We deal in this paper with some of specific problems raised by the grapheme to phoneme transcription of unknown words.

Our system works in several steps in the treatment of a text. We first compute all words which belong to our lexicon. Then, for each unknown word, we determine if it is a proper name, an acronym or a neologism. A morphological process estimates a syntaxic category for words tagged as neologisms.

We have several sets of transcription rules which are dependent on word categories. This tagging process allows us to determine which set of rules is correct according to the tag.

The use of statistic and rule-based methods in both the tagging and the grapheme to phoneme transcription processes, allows us to take into account several linguistic levels.

1. INTRODUCTION

La phonétisation automatique de documents n'est pas un problème nouveau, les systèmes existants donnent déjà de bons résultats. Toutefois c'est un domaine de recherche toujours actif permettant de tester et valider efficacement différentes techniques du Traitement Automatique du Langage Naturel (TALN).

Les premiers systèmes ont été développés selon une approche à base de règles (Prouts, 1980) codant les régularités de transcription phonétique des mots usuels du langage. S'il n'y a pas de bijection entre les transcriptions graphiques et phonétiques des mots de la langue française, il est néanmoins possible de dégager un nombre raisonnable de règles (environ un millier pour le français) ayant une couverture suffisamment large pour être intégré dans un système de synthèse de la parole. Les inévitables exceptions à ces règles sont alors codées, soit sous forme de liste, soit selon le formalisme même des règles.

Cette méthode donne de bons résultats, ceux-ci seraient meilleurs si l'on savait résoudre plusieurs problèmes résiduels tels le traitement des homographes hétérophones, des liaisons, la phonétisation des noms propres, des sigles ou encore des mots nouveaux. De plus, s'il est assez facile de produire des règles à partir d'une base d'exemples, il est plus complexe d'unifier les règles entre elles pour produire un ensemble minimal offrant la même couverture. La maintenance d'une telle base de règles est une opération compliquée et coûteuse en temps.

Ces raisons et la disponibilité de grandes bases de données d'exemples ont conduit la communauté scientifique à s'intéresser à des techniques d'apprentissage automatique. Ces techniques regroupent l'apprentissage symbolique, les réseaux neuronaux ou encore les modèles markoviens (voir (Yvon, 1994) pour une revue des différentes méthodes automatiques employées en conversion graphème-phonème).

Ces techniques, même si elles n'ont pas détrôné les systèmes à base de règles, permettent souvent de mieux prendre en compte la variabilité intrinsèque de la phonétisation d'une langue en particulier pour les mots non-usuels (noms propres, néologisme, mots étrangers, etc.).

Néanmoins, ces méthodes n'ont pas pour autant résolu tous les problèmes évoqués auparavant. Dans une autre veine on peut s'inspirer des phénomènes cognitifs impliqués dans la phonétisation. Citons pour exemple les systèmes fondés sur l'analogie (Dedina, 1991) et les systèmes intégrant plusieurs niveaux de connaissance. C'est dans cette voie que nous nous plaçons en faisant collaborer au sein de notre phonétiseur différents agents utilisant des informations graphiques, phonétiques, morphologiques, syntaxiques et sémantiques.

2. PRINCIPE DE PHONÉTISATION

L'écriture d'une base de règles de phonétisation permet de prendre en compte les

différents niveaux de connaissance impliqués dans un tel traitement. En effet, même s'il arrive qu'à une graphie ne soit associée qu'une seule forme phonétique, le plus souvent il y a ambiguïté quant à la transcription d'une lettre ou d'un groupe de lettres. Ces ambiguïtés peuvent être levées sur plusieurs niveaux :

•graphique (le "g" dans "garçon" et "geôle") ;
•syntaxique (la séquence "ent" dans "jugement" et "mangent") ;
•sémantique (le mot "fils" dans la phrase "La mère tisse des **fils** de laine à ses **fils**").

De plus le traitement de mots non-usuels (noms propres, sigles, mots nouveaux et d'origine étrangère) nécessite la mise en place de traitements particuliers pour répondre aux nouvelles ambiguïtés introduites.

Le système Griphon (Béchet, 1995) s'est donné comme objectif le traitement de textes journalistiques, il se doit donc de prendre en compte les problèmes posés par le traitement de ces mots non-usuels, et de manière générale de tous les mots inconnus. C'est un système utilisant des bases de règles de phonétisation. Ces règles sont contextuelles et peuvent être contraintes. L'intégration des différents niveaux linguistiques se fait en amont de cette phase de phonétisation par un étiquetage des mots composant la phrase traitée. C'est en fonction des étiquettes posées que telle ou telle base de règles sera employée.

3. ÉTAPE D'ÉTIQUETAGE

Le processus d'étiquetage consiste à affecter une catégorie syntaxique à chaque mot de la phrase à phonétiser. Pour cela nous utilisons le module d'étiquetage syntaxique développé au LIA (Spriet, 1995). Il repose sur une approche purement probabiliste. La stratégie consiste à générer le graphe orienté d'étiquettes candidates associées, puis à rechercher dans ce graphe le chemin de probabilité maximale. Un chemin est composé de transitions entre couples d'étiquettes, chaque transition s'accompagnant de l'émission d'un des mots du texte traité.

La probabilité de transiter peut être vue comme la probabilité conditionnelle de produire une étiquette en fonction d'un historique : le contexte gauche immédiat constitué des deux étiquettes précédentes (modèle tri-classe). À l'émission du mot est attachée la probabilité du mot sachant la classe prédite.

Bien que n'intervenant pas directement dans le calcul de la probabilité d'une étiquette, le contexte droit influe aussi dans le choix final de l'étiquette attribuée à un mot, influence qui s'exerce lors de la recherche du meilleur chemin.

Le corpus sur lequel nous avons travaillé est constitué des articles du Monde de 1992 et 1993, représentant un volume global de 39 millions de mots. Le dictionnaire utilisé, construit à partir de BDLEX2 (Pérennou, 1992) contient 201 000 formes fléchies. En complétant cet ensemble par la liste des noms propres apparus le plus fréquemment, nous obtenons sur un corpus de test une couverture statique de 97,4%. Aux mots inconnus restants est associée une étiquette spécifique MOTINC. Le jeu de 103 classes grammaticales que nous utilisons ici n'a pas été spécifiquement développé pour une application de phonétisation de textes. Cela explique le nombre qui peut paraître important de classes différentes n'ayant pas toutes d'utilité dans notre application. Néanmoins la mise au point d'un jeu optimal de classes en fonction d'une application donnée est un problème ouvert.

A l'issue de ce processus d'étiquetage nous sommes en mesure de savoir si le mot traité est un mot usuel (étiqueté par sa catégorie syntaxique), un nom propre (étiqueté XPREM pour les prénoms, XFAM pour les noms de famille et XSOC pour les noms de société) ou encore un mot inconnu (étiqueté MOTINC). En fonction du contexte et d'heuristiques simples sur la phrase, nous pouvons estimer si les mots étiquetés inconnus peuvent être aussi des noms propres, des sigles, ou des mots nouveaux. Nous allons décrire maintenant les traitements spécifiques associés à chacune de ces étiquettes.

4. LES SIGLES

Les sigles envahissent notre vocabulaire quotidien, leur prolifération est telle qu'on ne peut plus les ignorer dans les systèmes de TALN. La gestion d'un dictionnaire exhaustif de tous les sigles répertoriés avec leurs prononciations attestées se révèle insuffisante. En effet les tests effectués sur d'importants corpus journalistiques nous ont montré que l'univers des sigles, loin d'être clos, était en perpétuelle expansion.

Face à un sigle, on peut décider de le lire comme s'il s'agissait d'un mot ou de l'épeler lettre à lettre. Il ne s'agit pas d'un "ou" exclusif : en effet quelques sigles sont mi-lus, mi-épelés (V-DAT, CDROM) et d'autre admettent les deux modes d'oralisation (ONU, CES).

En ce qui concerne les processus de décision relatifs au choix d'oralisation, nous avons réalisé un module de décision lu/épelé à base de

règles inspirées des travaux sur le sujet (Plénat, 1993) (Boula, 1994). Les tests accomplis sur un corpus d'un millier de sigles issus du journal "Le Monde" ont donné un score de 96% de bonne identification du mode d'oralisation.

La décision du mode de prononciation du sigle (épelé ou lu) est dépendante du poids du sigle exprimé sous diverses unités. L'unité de compte pouvant être la lettre ou le phonème. Nous avons extrait des règles de décision en fonction de la structure des sigles. Ces structures sont représentées en consonnes (C), consonnes occlusives (CO), consonnes liquides (CL) ou voyelles (V).

Pour être lu, un sigle doit présenter au moins un doublet de type (CV). Ce qui implique que les sigles formés uniquement de consonnes ou uniquement de voyelles sont systématiquement épelés. Au delà de cette particularité commune à tous les sigles, il convient d'introduire des règles particulières selon le nombre de lettres composant le sigle.

Une fois le mode d'oralisation déterminé, nous appliquons dans le cas des sigles lus des règles de phonétisation spécifiques. En effet si les règles concernant les mots usuels suffisent dans la plupart des cas pour oraliser des sigles simples ("ONU", "ARC"), elles sont incomplètes pour d'autres sigles ("SGEN" qui se prononce (zgen) par exemple). Il a donc fallu mettre en évidence des règles de prononciation propres à l'oralisation des sigles. Ces règles ont été établies à partir de l'examen des 1000 sigles les plus courants extraits du corpus du journal "Le Monde".

5. LES NOMS PROPRES

L'oralisation des noms propres justifie à elle seule un traitement particulier, comme l'a mis en évidence le projet européen ONOMASTICA. La phonétisation des noms propres en français peut se décomposer en deux types différents de problèmes : d'une part le traitement des noms d'origine française, d'autre part ceux d'origine étrangère.

La phonétisation des noms d'origine française diffère de celle des mots usuels de manière significative. En effet des phénomènes d'archaïsme (**Isle**-sur-Sorgue prononcé /il/ et non pas /isl/), d'agglutination (**Montpellier** prononcé /mɔ̃pølje/ et non pas /mɔ̃tpølje/), de régionalisme (St **Gaudens** prononcé /godẽs/ dans le Sud-Ouest) (Yvon, 1993) perturbe l'application des règles d'oralisation classique.

D'un autre côté, la phonétisation des mots d'origine étrangère est grandement conditionnée par l'origine linguistique supposée du mot traité. C'est sur ce dernier point que nous avons axé nos efforts en réalisant un module statistique permettant de prédire une appartenance à un groupe linguistique en fonction de la graphie du mot à phonétiser.

À cet effet, nous avons extrait du journal "Le Monde" un corpus de 10 000 noms et prénoms. Ils ont ensuite été classés en fonction de certains traits communs caractéristiques de leurs prononciations. Cette classification a abouti à des ensembles disjoints de par leur consonance mais pas forcément leur graphie. Un ensemble de huit groupes linguistiques a ainsi été défini de façon empirique.

Nous avons construit, pour chacun de ces différents groupes, un modèle statistique tri-lettre à partir du corpus trié. Après apprentissage des différents modèles, il est possible de calculer pour un nom donné sa probabilité d'être associé à chaque groupe. Il est alors aisé de déterminer de manière automatique l'appartenance du nom examiné à un groupe de prononciation particulier.

Nous avons développé pour chaque groupe de prononciation un ensemble de règles censées représenter, non pas la phonétisation complète de la langue d'origine, mais plutôt la manière dont les particularités de la langue étudiée sont retranscrites usuellement en français.

Lors de l'utilisation du système les noms propres sont repérés grâce aux étiquettes syntaxiques posées. Une fois la classe de prononciation du nom déterminée à l'aide du modèle tri-lettre, le mot est phonétisé avec les règles de transcription phonétique relatives au groupe choisi.

6. LES MOTS INCONNUS

La dernière catégorie de mots traitée est celle des mots inconnus (étiqueté MOTINC) n'ayant pas été considéré comme des noms propres. Cette catégorie regroupe en particulier les néologismes absents des dictionnaires et dont les journalistes sont friands.

L'approche classique consiste à utiliser les règles de phonétisation des mots usuels du français. Cependant il apparaît primordial de leur affecter une catégorie syntaxique, et ce pour deux raisons: D'une part, comme nous l'avons déjà mentionné, l'ambiguïté de certaines graphies nécessite une connaissance syntaxique du mot traité. Ainsi dans le mot "redénationalisent", la graphie "ent" ne sera pas phonétisée de la même manière selon que ce mot sera considéré comme un verbe ou un nom.

D'autre part, les règles qui régissent les liaisons font souvent appel à des considérations syntaxiques (El-Bèze, 1990) (Grévisse, 1993).

Pour ces raisons, il nous a paru important d'incorporer dans notre système un module d'estimation syntaxique, appelé "devin", pour chaque nouveau mot rencontré. Nous avons choisi d'effectuer cette analyse morphologique à l'aide de méthodes statistiques rendues possibles grâce aux importants corpus en notre disposition.

La remarque préalable à cette étude est le fait que le suffixe d'un mot conditionne fortement sa catégorie syntaxique, surtout s'il s'agit des 21 classes correspondant aux substantifs, adjectifs, adverbes et verbes. Nous avons donc entraîné un modèle statistique sur les finales graphiques des mots de nos dictionnaires.

Le corpus d'apprentissage représente environ 100000 mots. Les tests effectués sur un ensemble de 28000 mots donnent les résultats suivants : dans seulement 4% des cas, l'étiquette de plus forte probabilité n'appartient pas à la liste des classes attendues. Le rang moyen des catégories possibles pour chaque mot est de 1,36 alors que le rang optimal est de 1,08.

Nous affectons ainsi aux mots nouveaux une catégorie syntaxique supposée avant de les phonétiser avec notre base de règles contraintes syntaxiquement prévues pour le traitement des mots usuels.

7. CONCLUSION

L'utilisation conjointe de techniques statistiques et à base de règles nous a permis de prendre en compte différents niveaux linguistiques : l'étiqueteur syntaxique propose une catégorie syntaxique pour chaque mot ; le "devin" morphologique estime une catégorie syntaxique pour les mots inconnus ; le module de traitement des noms propres permet d'affecter une origine linguistique à chaque nom et enfin le module de décision lu/épelé décide du mode d'oralisation des sigles. Cette phase d'analyse du texte permet de choisir parmi les bases de règles de phonétisation celle qui s'applique à la catégorie de mot calculée.

Cette approche nous évite la gestion d'une base de règles énorme devant prendre en compte tous les cas possibles de phonétisation. Chaque module ainsi créé est indépendant et possède son propre phonétiseur. Il est donc possible de modifier facilement le processus de phonétisation (en utilisant, par exemple, des techniques d'auto-apprentissage pour le traitement des noms propres) sans remettre en cause la phase d'analyse et d'étiquetage du texte.

De plus, ces techniques s'appliquent particulièrement aux mots hors-dictionnaire (noms propres, sigles, mots nouveaux) en estimant pour chacun d'eux une catégorie qui conditionnera leur phonétisation.

8. BIBLIOGRAPHIE

Béchet F., Derderian D., El-Bèze M. (1995) Conversion graphème-phonème automatique le système GRIPHON ; *IA 95, Génie Linguistique, 15èmes Journées Internationales ; Montpellier.*

Boula de Mareuil P. (1994) Vers une phonémisation automatique des sigles ; *actes des XXèmes Journées d'Étude sur la Parole, Trégastel,* pp. 95-100.

Dedina M.J., Nusbaum H.C. (1991) Pronounce : a program for pronounciation by analogy ; *Computer Speech and Langage* 5:55-64.

El-Bèze M. (1990) Choix d'unités appropriées et introduction de connaissances dans des modèles probabilistes pour la reconnaissance automatique de la parole ; *Thèse de Doctorat, Université Paris VII.*

Grevisse M. (1993) Le Bon Usage, grammaire française, refondue par A. Goosse, Duculot.

Pérennou G., de Calmès M., Ferrané I., Pécatte J.M. (1992) Le projet BDLEX de base de données lexicales du français écrit et parlé ; *actes du Séminaire Lexique, IRIT-UPS Toulouse*, pp. 41-56.

Plénat M. (1993) Observation sur le mot minimal français - L'oralisation des sigles ; *De Natura Sonorum : Essais de phonologie*, Presses Universitaires de Vincennes, pp. 143-172.

Prouts B. (1980) Contribution à la synthèse de la parole à partir de texte, transcription graphème-phonème en temps réel sur micro-processeur ; *Thèse de Docteur Ingénieur, Université de Paris Sud* (Paris XI; Orsay).

Spriet T., El-Bèze M. (1995) *Étiquetage probabiliste et contraintes syntaxiques* ; TALN'95, Marseille.

Yvon F. (1993) A tidy rule-based grapheme to phoneme transcriber for ONOMASTICA ; *actes du 1er colloque ONOMASTICA*, pp 47-56, Londres.

Yvon F. (1994) Self learning techniques for grapheme-to-phoneme conversion ; *actes du 2em colloque ONOMASTICA*, pp 25-38, Londres.

LA PHONÉTISATION D'UN LEXIQUE DE RÉFÉRENCE DU FRANÇAIS : ÉMERGENCE DE SYSTÈMES HYBRIDES RÈGLES/LEXIQUES

Rabia BELRHALI et Véronique AUBERGÉ
Institut de la Communication Parlée
INPG/Université Stendhal, BP 25, 38040 Grenoble Cedex 9, FRANCE
Tél : 76 82 41 17 Fax : 76 82 43 35 - email : (belrhali,auberge)@icp.grenet.fr

ABSTRACT

This study aims at describing the relations between the orthographic code and the phonetic code in a French letter-to-phone system. The *graphone*, minimal unit of logical operating, inferior to the grapheme, emerges from this analysis. Our concern is to explain the linguistic use of this unit. It is shown that the general lexicon of language includes sub-lexicons, defined by similar functional and etymological characteristics (*e.g.* same output language, same usage). These sub-lexicons cannot be related to an ideographic processing of language: they are the image of the phonological transfer from the output phonographic system to phonographic French system.

1. INTRODUCTION

Les premiers travaux traitant des relations entre le code écrit et le code oral avaient pour objectifs (1) la publication de dictionnaires de prononciation (Féline 1851) (2) l'élaboration de grammaires qui enseignent à « bien lire, à parler purement, et à écrire correctement » (Girault-Duvivier Ch. P. 1818).

L'objet de cette étude est de faire émerger l'organisation linguistique des relations entre le code orthographique et le code phonétique, en s'attachant, dans une première étape décrite ici, au passage de l'orthographe vers la phonétique. L'outil formel de cette description est un langage de grammaire-lexique TOPH (Aubergé 1991) qui autorise le traitement automatique, pour une validation *a posteriori* des systèmes mis en évidence. La base de cette étude est l'analyse systématique d'un large corpus : un lexique du français constitué des formes canoniques du lexique électronique *Le 60 000* de l'ICP, associées aux références phonétiques du *Petit Robert 1* (1990) puis, dans un deuxième temps, les formes fléchies dérivées des formes canoniques du *60 000*. Nous avons appliqué une méthodologie qui consiste (1) dans une approche inductive : à appliquer une analyse logique optimale au lexique, sans *a priori* linguistique, sur les correspondances lexie orthographique/lexie phonétique (« lexie », i.e. unité de lexique, sera préférée à « mot » qui reste flou), afin de mettre en évidence le fonctionnement minimal des unités liées dans cette correspondance (2) dans une approche déductive : à vérifier systématiquement la validité linguistique des unités et des systèmes qui émergent de l'analyse inductive.

L'exploration logique de ce lexique général de la langue amène à définir une unité de fonctionnement logique composite du graphème, le *graphone*, dont la validité linguistique sera démontrée.

Cette étude montre que (1) le lexique général de la langue peut se décrire par un ensemble de règles de réécriture graphoniques sous contraintes contextuelles (2) des sous-lexiques sont isolés du lexique général et se décrivent par des sous-systèmes de réécriture : ils sont caractérisés par un fonctionnement logique spécifique corrélé à une similitude de traits diachroniques ou synchroniques (e.g. emprunt, usage).

2. MÉTHODOLOGIE

La méthodologie sur laquelle repose cette étude est déterminante. Elle est en tout cas originale en regard des nombreux travaux sur la phonétisation automatique.

Il s'agit schématiquement (1) de mener un apprentissage par expertise (mais selon des contraintes strictement logiques, i.e. sans a priori linguistique) sur un corpus représentatif, (2) de valider l'apprentissage correct du corpus, (3) de vérifier la capacité de généralisation. Ce processus limité à (1) et (2) est évidemment tautologique. Mais si le corpus est un échantillon de très large couverture, l'étape (3) se ramène à l'étape (2) et la généralisation est implicite. C'est pourquoi nous avons choisi un corpus suffisamment large pour décrire quasi-exhaustivement les formes de la langue. Ainsi, la généralisation d'une description résultant d'un corpus équivalent au Petit Robert est immédiate : les lexies absentes sont soit de fréquence d'usage général faible, soit des

néologismes qui répondent aux mécanismes décrits dans le lexique d'apprentissage.

Cette description par apprentissage s'est elle-même organisée en trois étapes inclusives, selon une méthode de *bootstrapping*. La première étape a consisté à établir une grammaire de base des transcriptions phonographiques (Aubergé et al., 1987 ; Belrhali et al, 1992). À partir de ce filtre primitif appliqué sur le lexique des 52 381 formes canoniques, nous avons déduit une grammaire-lexique qui décrit les mécanismes de transcription selon des contraintes d'optimisation logique. Afin de cerner exhaustivement toutes les réalisations, chaque lettre a fait l'objet d'une recherche systématique selon sa position dans l'entrée lexicale. Cette grammaire est ensuite appliquée aux lexiques des formes fléchies déduites des formes canoniques. Dans cette seconde étape, sont annotés les graphèmes flexionnels transcrits par la grammaire-lexique des formes canoniques, et sont construites les règles spécifiques aux flexions (Sannier, 1995). La troisième étape a pour but de décrire le fonctionnement des formes en contexte textuel, c'est-à-dire de traiter des phénomènes de liaison. En restant dans la même optique méthodologique, la liaison a été observée et décrite à partir de ce corpus. Ce travail, décrit dans Ahmad (1993), a débouché sur une typologie statistique des liaisons en contexte syntaxique et morphologique dans le corpus BDPHO (Boë et Tubach, 1992).

La grammaire substrat est constituée des règles primitives correspondant aux réécritures régulières de base (220 règles) et de la liste des principaux lexiques (12 lexiques). Le lexique des formes canoniques du français est *Le 60 000* de l'ICP, à partir duquel seront calculées les formes fléchies. L'unité maximale du lexique renvoie à la lexie. Notre objectif n'était pas de décrire les variantes phonétiques de chaque entrée lexicale. Nous avons donc été confrontées au choix d'une référence, d'une norme de prononciation. La définition d'une norme repose sur des critères socioculturels dont la mise en œuvre n'aboutit pas à une solution univoque. La référence phonétique que nous avons retenue est celle du Petit Robert 1, dictionnaire de langue, qui associe une forme phonétique à chaque entrée, à partir de l'expertise d'une phonéticienne A. Boumendil-Lucot dans l'édition nous a servi de référence, celle de 1990. La cohérence de ce choix peut donc, au cas par cas, être contestable. Cependant, l'échantillon représenté par ces données est suffisamment large pour que la cohérence globale puisse clairement apparaître.

3. GRAPHÈME *VS.* GRAPHONE

L'établissement des règles a été mené selon des contraintes systémiques : opposition/similitude des unités de correspondance. Cela nous a amenées à reconsidérer le choix de l'unité minimale du système de correspondances. Le graphème utilisé par la plupart des auteurs (Catach, 1980 ; Thimonnier 1967 ; Mounin, 1974 ; Béchade 1992, Anis 1988) était un bon candidat, généralement décrit comme unité minimale. Cependant, la définition du graphème est variable selon le courant dans lequel il s'inscrit et surtout peu précise sur le plan fonctionnel.

Les définitions proposées ne rendent pas précisément compte des unités qui ont pu émerger de notre étude. De l'analyse logique s'est détachée le graphone, unité minimale irréductible, composite du graphème : celle qui correspond à la sous-chaîne réécrite dans la grammaire TOPH. Le graphème est ensuite défini, dans cette étude, à partir du graphone.

Nous ne donnons pas ici de définition formelle du graphone mais nous allons l'illustrer par un exemple, celui de la chaîne "qu" généralement donnée comme graphème. Diachroniquement, si l'on s'intéresse à la prononciation [kw], on note que cette prononciation était représentée dans l'orthographe latine par la graphie "qu". Le trait labial [w] avait disparu jusqu'à la fin du XII^{e} siècle et a été réintroduit en français moderne dans des mots savants issus du latin (tels que « quadrilatère »). Il faut remarquer que de nouvelles prononciations sont apparues, notamment la prononciation [kɥ] devant "i", pour « noter une tendance savante à la prononciation » (Catach, 1995).

Dans notre corpus, "qu" apparaît dans 3444 lexies, dont 3303 avec la prononciation [k], 109 avec la prononciation [kw] et enfin 32 référencés avec la prononciation [kɥ]. Pour composer les règles de phonétisation relatives au graphème "qu", nous avons défini une fenêtre d'observation (chaîne à transcrire et contextes gauche et droit) appelée focus.

Dans le cas de "qu-" en position initiale de la lexie « quatre », la prononciation référencée est [k]. Le focus, c'est-à-dire la fenêtre d'observation de la règle (contextes et graphone), sera la chaîne "qu" concaténée à la voyelle suivante, le fonctionnement des unités graphiques sera représenté de la manière suivante :

+q+ —> [k], (q)+u+ —> []

(*i.e.* "u" ne se prononce pas)

Ce fonctionnement permet de mettre en évidence deux graphones : "q", "u", la non réalisation du graphone "u" reflétant donc la disparition de la labialité. La voyelle suivante suit une transcription indépendante de "u" ou de "qu". La transcription de "u" étant strictement liée à "q", "qu" est un graphème.

Il est évident que les graphones ainsi définis ne sont pas, dans tous les cas, équivalents à une lettre graphique. Par exemple, le graphème "gn" qui trouve sa réalisation phonétique en [ŋ] dans « agneau », est un graphone. Mais quand sa réalisation est [gn] (comme dans « magnum ») il s'agit de deux graphones : le graphone "g" prononcé [g] et le graphone "n" prononcé [n] qui ne sont pas liés ici en graphème.

4. LES SOUS-SYSTÈMES

Des ensembles de lexies ont été mises en évidence. Ils ne doivent pas être assimilés à un mécanisme de nature idéographique, comme on a pu parfois le proposer pour le français mais au contraire être reliés à des sous-systèmes de correspondances graphoniques régulières, non pas globalement propres au lexique « général » du français, mais spécifiques aux « sous-langages » définis par les lexiques. Ce sont des régularités locales aux lexiques donnés : les langues sources, auxquelles le français a emprunté des lexies, fonctionnent selon des mécanismes phonographiques transférés sur le français; la description de ces phénomènes dans TOPH rend compte de ces transferts phonologiques.Ainsi, l'une des règles TOPH qui fait appel au lexique des emprunts anglais contemporains s'énonce :

("#"+ "Lexique-emprunts-anglais-XXe") +ng+ ("#") -->[ŋ]
et non pas

("#") +parking+ ("#") --> [paʀkiŋ].

De la même façon qu'il est montré que les règles phonographiques de TOPH sont généralisables, il peut être proposé une généralisation des sous-lexiques : il suffirait d'étiqueter (ou de détecter automatiquement, par exemple par des analyses en n grammes spécifiques du langage) l'origine linguistique d'une lexie nouvellement empruntée et de lui appliquer les règles définies dans TOPH pour les lexies de mêmes caractéristiques.

Un lexique désigne donc une même particularité de réécriture, bien souvent expliquée par des faits semblables (mots d'emprunt de même origine, de même usage...). Il arrive que certaines lexies n'appartiennent à aucune grande famille et se trouvent isolées dans des lexiques spécifiques.

Ainsi, nous avons établi la liste de tous ces lexiques et vérifié que les mots de chaque lexique partagent une même classe historique par une vérification étymologique systématique (Belrhali, 1995).

La grammaire-lexique TOPH du français comporte 24 règles qui activent 24 lexiques décrivant 4 442 lexies. Les lexies d'un lexique TOPH partagent le même fonctionnement graphonique.

Ci-dessous est donné en illustration un exemple de lexique :

• Le lexique des entrées dans lesquelles "b" suivi de "s" se prononce [b] (et non pas [p]), se compose de 8 entrées lexicales issues du latin :

...

subsidence : 1875 ; avec le sens « sédiment, dépôt », 1557; latin « *subsidentia* » ;

subsidiaire : XVIe ; avec le sens « de renfort », 1352; latin « *subsidiarius* » qui signifie « de réserve » (de troupes) ;

subsidiairement : XVIe ; à partir de « subsidiaire » ;

subsistance : XVIIe; d'un édifice, 1514; de « subsister » ; ...

et de 2 lexies introduites dans le lexique français par le biais de l'anglais :

lambswool : milieu XX^{e} ; de l'anglais.

bobsleigh : 1889, mot anglais

5. CONCLUSION

Cette étude s'est donnée (1) pour but de montrer qu'une description linguistique systémique peut émerger d'une analyse logique, c'est-à-dire que les systèmes linguistiques de la phonétisation du français répondent avant tout à des critères d'optimalité qui trouvent leur justification en synchronie comme en diachronie (2) pour moyen une méthodologie systématique basée sur un formalisme (TOPH) et sur un corpus représentatif du lexique de la langue (*Le 60 000*/Le Petit Robert).

La grammaire-lexiques déduite de l'analyse de ce corpus avoisine un taux de correction proche de 100% sur ce même corpus, la couverture linguistique de la grammaire se ramenant à l'exhaustivité du corpus. On peut supposer que ce lexique couvre l'ensemble des phénomènes de réécriture qui sous-tendent la généralisation de la phonétisation à des néologismes qui n'appartiennent pas à des

lexiques spécialisés (ou à des lexiques de sigles, acronymes ou noms propres).

L'un des points essentiels de ce travail est l'inclusion, dans le lexique général, de sous-lexiques décrits pas des sous-systèmes. Les sous-lexiques ne dénotent pas un fonctionnement idéographique du français : ils sont une trace du transfert dans le français moderne des systèmes phonographiques des langues sources. Nous avançons qu'il ne s'agit pas là de phénomènes de correspondances globales entre les mots orthographiques et leur prononciation mais au contraire de la rémanence linguistique des systèmes phonologiques empruntés, généralement de type phonographiques. Ce mécanisme linguistique peut-il être corrélé, sur le plan des processus cognitif en jeu dans la lecture, à une identification de la langue-source : lorsque la structure graphique d'un mot nouveau est identifiable par le lecteur à l'un des ses sous-lexiques, lui appliquera-t-il les règles qui décrivent ce sous-lexique ?

Une autre notion importante est celle de graphone. Il nous a semblé intéressant de montrer, qu'à partir d'une démarche analytique formelle sur un état synchronique donné d'un lexique de la langue, ont émergé des unités fonctionnelles dont on montre la validité diachronique et à partir desquelles peuvent être construits les graphèmes : il s'agit maintenant d'en comprendre la valeur linguistique : le graphone est-il une unité de description des variantes de graphèmes ou bien une unité à part entière ? C'est donc une description basée sur une économie linguistique et formelle qui a été ainsi mise en évidence. On peut se demander si elle doit être reliée à une technologie cognitive des relations entre le code écrit et les unités de l'oral, aussi bien dans l'apprentissage de la lecture en ontogenèse qu'en phylogenèse.

Ce travail qui s'est donné pour objectif principal l'étude linguistique des phénomènes de phonétisation, trouve une application en synthèse de la parole (système COMPOST de l'ICP, Bailly et Tran, 1989). Cette grammaire TOPH a été écrite de façon à décrire les correspondances dans une relation nécessaire et suffisante. Ainsi, sa grammaire inverse d'orthographisation PHOT (Ghneim et Aubergé, 1995) devrait permettre de réaliser une description bidirectionnelle des relations entre le code écrit et le code phonétique.

6. BIBLIOGRAPHIE

Ahmad M. (1993) Vingt heures de français parlé : aspect phonétique de la liaison. *Thèse de troisième cycle, Université Stendhal, Grenoble.*

Anis J. (1988) *L'écriture.* Théories et descriptions. Éd. Universitaires, Paris.

Aubergé V., Contini M., Maret D., Schnabel B. & Zinglé H. (1987). Un outil de phonétisation multilangue. *XI Int. Cong. of Phon. Sci. Se 73-2.*

Aubergé V. (1991) La synthèse de la parole : « des régles aux lexiques ». *Thèse de troisième cycle, Université Pierre Mendès France, Grenoble.*

Bailly G. & Tran A. (1989) Compost : a Rule-Compiler for Speech Synthesis. *Eurospeech.*

Béchade H-D. (1992) *Phonétique et morphologie du français moderne et contemporain.* Presses Universitaires de France, Fondamental, Paris.

Belrhali R., Aubergé V. & Boë L-J. (1992) From Lexicon to Rules: Toward a Descriptive Method of French Text-to-Phonetics Transcription. *Internat.Conf. on Spoken Language Processing, Banff, Alberta,Canada, pp. 1183-1186.*

Belrhali R. (1995) Émergence et systémique linguistique : phonétisation automatique d'un lexique général du français. *Thèse de l'Université Stendhal, Grenoble.*

Boë L.-J. & Tubach J.P. (1992) Une base de données lexicale orthographique-phonétique du français parlé. *Cahiers de Grammaire, 17, pp. 3-23, Toulouse.*

Catach N. (1980) *L'orthographe française. Traité théorique et pratique.* Ed. Nathan, Paris.

Catach N. (1995) *Dictionnaire historique de l'orthographe française.* Sous la direction de Nina Catach, Ed. Larousse, Trésors du Français, Paris.

Féline A. (1851) *Dictionnaire de la prononciation.* Firmin Didot Frères Éditeurs, Paris.

Ghneim N. & Aubergé V. (1995) Optimising the French Letter-to-Phone Grammar TOPH With a View to Phonographic Spelling Correction. *ROCLING, Séoul.*

Girault-Duvivier Ch. P. (1818) *Grammaire des Grammaires ou Analyse raisonnée des meilleurs traités sur la langue française.* Troisième Édition, Janet et Cotelle Libraires, Paris.

Le Petit Robert 1 (1990) *Dictionnaires Le Robert.*

Mounin G. (1974) *Dictionnaire de la linguistique.* Quadrige, Presses Universitaires de France, Paris.

Sannier F (1995) Les règles flexionnelles dans la grammaire TOPH de phonétisation du français. *Rapport de DESS en Industries de la Langues de l'Université de Metz.*

Thimonnier R. (1967) *Le système graphique du français.* Ed. Plon, Paris.

UTILISATION D'UN SYSTÈME DE RECONNAISSANCE DE LA PAROLE POUR ACCÉDER À W3

Eric Thiébaut, Jean-François Mari, Jean-Paul Haton, Yifan Gong, Dominique Fohr

CRIN-CNRS & INRIA Lorraine, Bâtiment Loria, B.P. 239, 54506 Vandoeuvre-lès-Nancy

Tél.: 83.59.20.00 - Fax: 83.41.30.79 - e-mail: {jfmari,jph,gong,fohr}@loria.f

ABSTRACT

This paper presents a speech understanding system capable to give access to W3. To achieve this task, the system implements 3 process. The first one, called Vinics, is a continuous speech recognition system for artificial langage. The second one carries out the interpretation of the user requests and builds various grammars to adapt the language to the serveur being accessed, the last one, derived from Mosaic gives access to W3. Preliminary results and examples are given in section 7.

1. INTRODUCTION

Internet est un ensemble mondial de réseaux interconnectés. Divers protocoles de communications cohabitent sur celui-ci, notamment World Wide Web plus connu sous ses abréviations WWW ou encore W3.

Divers logiciels permettent d'exploiter ce dernier: citons entre autres Mosaic et Netscape. Ces logiciels permettent à l'utilisateur de visualiser des documents à l'aide d'une interface graphique. Par un simple click de sa souris sur une phrase soulignée ou une image, l'utilisateur peut alors visualiser un nouveau document.

Cependant, l'interface graphique n'est pas le moyen le plus naturel de dialogue pour l'homme et il est séduisant d'envisager un dialogue oral pour accéder à W3. Un des avantages de la parole est de pouvoir désigner des objets non présents à l'écran, ce qui est impossible de réaliser à l'aide de la souris.

Cet article décrit un système permettant de simplifier l'accès aux documents mutimédias du réseau W3 en utilisant le système Vinics (Gong, 1994), système de reconnaissance de parole continue développé par l'équipe RFIA du CRIN.

2. PRINCIPE DE FONCTIONNEMENT DE W3

Pour l'adressage des documents, W3 utilise une technique dite hypertexte permettant de référencer toutes sortes de documents disponibles sur Internet. Elle est fondée sur le concept d'URL (Uniform Resource Locator) qui spécifie, entre autre, le nom de machine du serveur possédant le document ainsi que le nom du fichier.

Un lien hypertexte, ou hyper-lien, est formé par une ancre et par l'adresse du document ciblé. Une ancre peut être un mot (ou un groupe de mots) ou une image mis en évidence (caractères gras, couleurs, encadrement...) dans le document.

3. INTÉGRATION DE MOSAIC ET VINICS

Comme tout système de reconnaissance de parole continue, Vinics nécessite l'apport d'une grammaire. Dans le contexte de notre application, cette grammaire décrit un ensemble de requêtes que l'utilisateur pourra prononcer.

3.1. Spécification des requêtes

On distingue deux types de requêtes : commandes et demandes de pages.

- Les commandes sont elles-même divisées en deux groupes :
  - les commandes à destination de Mosaic, comme «descendre», «monter», «ouvrir une nouvelle fenêtre», etc. ;
  - les requêtes de changement de mode. Celles-ci sont destinées à changer le mode dans lequel l'utilisateur souhaite se placer (cf. paragraphe 3.2.).
- Les demandes de pages, qui sont toutes construites sur la même forme:
  - une grammaire statique incomplète qui décrit les formules de politesse pour exprimer une demande ;
  - une sous-grammaire dynamique qui comporte, en fonction du mode de fonctionnement, soit un mot, soit une suite de mots qui sont des points d'ancrage des pages HTML analysées.

Ces deux grammaires se complètent. En effet, la sous-grammaire dynamique s'insère

dans la grammaire statique par l'intermédiaire d'un non-terminal prédéfini.

3.2. Interaction homme-machine durant une session W3

Plusieurs types de personnes utilisent W3 et pour des buts différents. Il s'avère donc nécessaire d'interpréter différemment leurs requêtes; pour cela nous avons introduit différents modes:

- Certaines personnes ne connaissent pas le contenu des pages, ou la localisation de la page qui correspond à leur recherche. Ainsi, le plus simple pour celles-ci est de donner un mot-clé indiquant leur recherche: c'est le mode «débutant».
- Un deuxième type de personnes naviguant fréquemment sur le réseau et par conséquent connaissant bien quel point d'ancrage est lié au document recherché, peuvent simplement faire une requête comportant exactement la suite de mots correspondant à ce point d'ancrage : c'est le mode «expert».
- D'autres personnes, bien que non expertes, souhaitent avoir plus de possibilités que n'en propose le mode débutant. Par exemple, elles veulent faire une recherche par rapport à une ancienne page déjà analysée. Le mode «choix» permet ce type de recherche, il permet aussi de choisir entre plusieurs adresses de page (URLs) lorsque le mot-clé prononcé en réfère plus d'une.

4. MODIFICATION DE MOSAIC

Lorsque l'utilisateur demande un document, Mosaic envoie une requête au serveur le possédant. Si ce document est une page HTML, il est possible de récupérer les phrases des points d'ancrage. La recherche se poursuit avec les points d'ancrage de ces nouvelles pages dans une limite de profondeur donnée (actuellement 2). Ainsi, un utilisateur pourra accéder oralement à des pages sans voir leurs points d'ancrage, ce qui constitue une possibilité supplémentaire par rapport à une désignation par souris.

5. MODIFICATION DE VINICS

Vinics est un ensemble de processus séparés permettant la reconnaissance de phrases simples. Cet ensemble de processus regroupe l'analyse acoustique, la reconnaissance, la compilation et la génération automatique des grammaires. Il existe une liaison de communication entre le système de reconnaissance et l'application permettant une nouvelle compilation de grammaire à chaque nouvelle page.

Vinics effectue une reconnaissance de phrases en commençant par une reconnaissance phonétique dont les paramètres dépendent du locuteur. La recherche lexicale est conduite par un analyseur syntaxique qui valide des hypothèses de mots à l'aide des transcriptions en phonèmes stockées dans le dictionnaire BDLEX (Pérennou, 1992).

La figure 2 décrit le fonctionnement de VINICS et son interaction avec Mosaic:

- Enregistrement d'un utilisateur :
  l'enregistrement d'un utilisateur nécessite que la phase d'apprentissage ait déjà été réalisée. Ce service permet l'initialisation de l'ensemble des paramètres propres à l'utilisateur. Il comporte donc une base de données contenant, pour chaque utilisateur potentiel, ses paramètres caractéristiques, comme par exemple les modèles de phonèmes.
- Génération et compilation automatique d'une grammaire :
  la reconnaissance d'une phrase exige la présence d'un réseau de reconnaissance (l'élément N° 5 dans la figure 2.). Ce réseau provient de la compilation de la grammaire (4), possédant une partie statique et une partie dynamique (3). Cette deuxième partie est générée automatiquement à partir d'un corpus de phrases (2), lui-même élaboré à partir de l'ensemble des liens (1) que Mosaic a découvert lors de sa recherche dans W3.
- Reconnaissance d'une phrase :
  la reconnaissance d'une phrase est menée à partir du signal de parole (6) représentant cette phrase. Elle utilise le dernier réseau généré lors de la compilation ou du chargement d'une grammaire.

Le résultat renvoyé est une suite de phrases classées par ordre décroissant de taux de reconnaissance.Une fois les phrases reconnues (7), celles-ci sont interprétées de manière à livrer à Mosaic une commande (8)

6. INTERPRÉTATION D'UNE REQUÊTE PROVENANT DE VINICS

Les requêtes proviennent de Vinics sous forme de phrases. Ces requêtes sont de deux

types (cf. 3.1). Dans une première étape, le début de phrase est analysé pour déterminer ce type. S'il s'agit d'une commande, celle-ci est alors soit exécutée dans le cas d'un changement de mode, soit transférée vers Mosaic.

Si la requête est une demande de page, elle est alors affinée de manière à obtenir la phrase clé qui va permettre de rechercher à travers le graphe des données l'URL correspondant au document souhaité.

Suivant le mode dans lequel se trouve l'application, cette phrase clé est une série de mots ou un mot seul, dont l'interprétation est fonction de ce mode. Nous allons donc, pour chacun de ces modes, donner un exemple d'interprétation de requête.

7. EXEMPLES D'UTILISATION

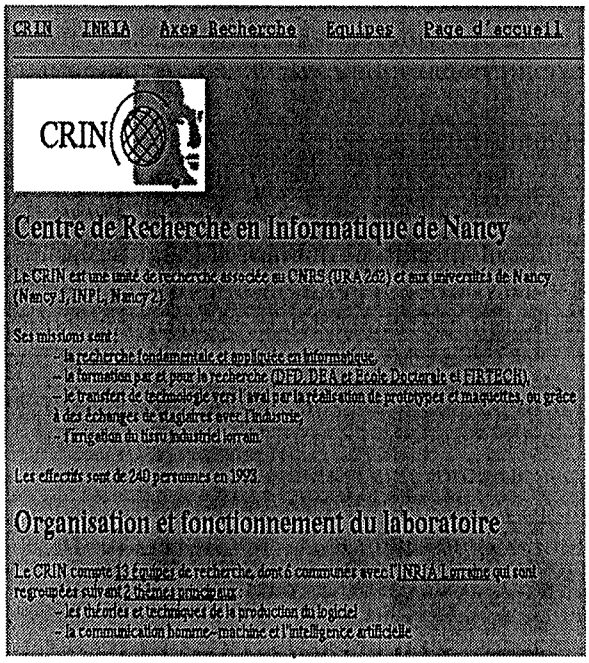


Figure 1 : Une page du serveur du CRIN

On considère une consultation au serveur du CRIN. La page donnée dans la figure 1 est située juste sous la page d'accueil.

Mode débutant

L'utilisateur a sous les yeux la page ci-dessus. S'il prononce la phrase «je voudrais les axes», c'est le point d'ancrage Axes Recherche qui est choisi car c'est la première occurrence de ce point d'ancrage qui est retrouvé dans le texte HTML. Si ce mot avait été absent, il aurait été recherché dans les pages connexes (cf. 4).

Mode expert

Dés la page d'accueil, dans ce mode, l'utilisateur peut prononcer «je voudrais la recherche fondamentale et appliquée». L'effet est le même que s'il avait désigné ce point d'ancrage dans la page ci-dessus.

Mode choix

Après avoir prononcé la phrase «je voudrais la recherche», le système analyse toutes les occurrences du mot «recherche» depuis la page d'accueil et les pages déjà accédées. Il

propose ensuite les choix suivants :
- Recherche fondamentale et appliquée ?
- Axes de recherche ?
- ...

8. CONCLUSIONS

Nous venons de décrire un système de consultation orale de W3. Le langage des requêtes est rigide et ressemble à un langage de commande par mots clés. Le principal problème provient de la génération dynamique des grammaires à la suite des accès aux pages et pendant la reconnaissance. Pourtant il s'agit de grammaires simples de langages artificiels très contraints.

A l'heure actuelle nous ne pouvons pas désigner autre chose que des groupes de mots en les prononçant. La désignation de plusieurs photos n'est pas possible. Nous sommes donc encore loin d'un système de demande orale en langue naturelle de renseignements. Ce système constitue néanmoins un bon terrain d'expériences pour mettre à jour et résoudre les problèmes de demande de renseignements dans les systèmes de demain et d'après demain.

9. BIBLIOGRAPHIE

Gong Y., Haton J.P. (1994) "Stochastic Trajectory Modeling for Speech Recognition", *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, 1994, 57-60

Pérennou G., Cotto D., De Calmes M., Ferrané I., Pecatte J.M. (1992) "Le projet BDLEX de base de données lexicale du français écrit et parlé", *Séminaire Lexique*, Toulouse

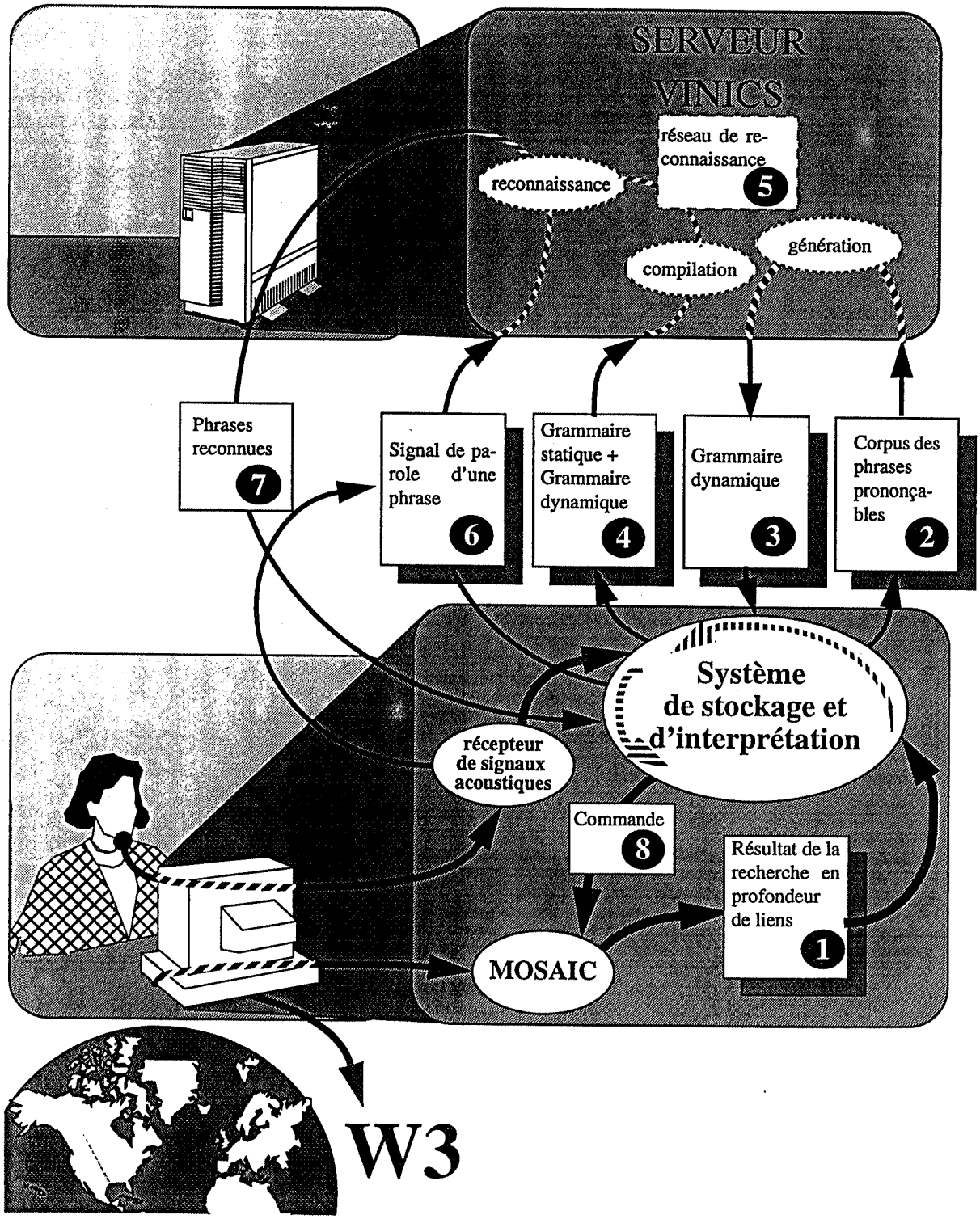


Figure 2 : Schéma de dépendance inter-processus

VERS L'ORTHOGRAPHISATION DU FRANÇAIS : DE TOPH À PHOT

Nada GHNEIM, Véronique AUBERGÉ
Institut de la Communication Parlée – INPG/Université Stendhal – BP 25 – 38040 Grenoble Cedex 9, FRANCE
TÉL.: 76 82 41 17 – FAX: 76 82 43 35 – e-mail: (ghneim,auberge)@icp.grenet.fr

ABSTRACT

If letter-to-phone processing in French has been well studied concerning linguistics and application field, the phone-to-letter processing has been less described, and mainly by context-free relations. This paper shows how a contextual phone-to-letter grammar-lexicon can be derived from a TOPH letter-to-phone description in a bi-directional approach. The development environment, which is crucial to establish such kind of knowledge, is presented. The problems raised by the inversion of a TOPH rule into a PHOT rule are exposed : the contexts of each TOPH rule, which are expressed in an orthographic form, must be transliterated into a phonetic form by applying the TOPH grammar, in order to generate each PHOT rule.

1. INTRODUCTION

La phonétisation automatique du français, c'est-à-dire la réécriture de la forme orthographique en une forme phonétique, est un sujet déjà longuement traité, certainement parce que les motivations sont multiples : description de normes de prononciation, enseignement du français, synthèse ou reconnaissance vocale, correction de l'orthographe ou plus généralement description linguistique. Les systèmes de phonétisation des langues alphabétiques sont divisés en deux groupes, selon qu'ils sont basés sur des connaissances explicites (lexiques de correspondances, grammaires de transcription) ou implicites (réseaux connexionnistes). Pour l'anglais, la capacité des systèmes neuronaux à modéliser la transcription graphèmes-sons a été clairement montrée inférieure aux systèmes mettant en œuvre des descriptions linguistiques (Ainsworth 90).

L'orthographisation, qui est le processus inverse à la phonétisation, a été beaucoup moins traité, principalement en vue de la correction automatique (Laporte 89; Strube de Lima 90; Véronis 93) ou de la reconnaissance vocale (Néel 86; Lacheret-Dujour 90, Hunnicutt 93). Alors que les systèmes de phonétisation à connaissances explicites sont généralement fondés sur des grammaires de règles terminales contextuelles (Divay 77; Catach 84) la plupart des systèmes d'orthographisation mettent généralement en œuvre des correspondances sons-graphies sans contraintes contextuelles, la cohérence des réécritures à travers le mot étant assuré par un filtrage lexical ultérieur.

Le but de ce travail est de montrer, pour le français, que si les connaissances phonographiques décrites dans la relation des graphies aux sons répondent à certaines contraintes méthodologiques à la fois formelles et linguistiques, elles peuvent devenir nécessaires et suffisantes. La relation des sons aux graphies se calcule alors par inversion de la relation des graphies aux sons avec une approche bidirectionnelle. Ainsi l'inversion d'une grammaire-lexique du français en langage TOPH (Aubergé 91) conduira à la description des relations sons-graphies formalisées en langage PHOT.

Nous montrerons l'importance et l'usage méthodologique de l'environnement développé autour du phonétiseur TOPH, puis la méthode utilisée et les difficultés calculatoires rencontrées dans l'inversion de TOPH à PHOT.

2. L'ENVIRONNEMENT TOPH/PHOT

2.1 La syntaxe

TOPH est à la fois un langage et un environnement informatique qui permet au linguiste de formaliser son raisonnement (de phonétisation) sous la forme d'une grammaire déterministe et ordonnée de règles terminales de réécriture contextuelle.

Une grammaire en langage TOPH [PHOT] est constituée d'une partie de déclaration d'ensembles et de lexiques définis sur le vocabulaire d'entrée (V=graphies ou S=sons) et d'une partie règles, dans laquelle sont appelés les ensembles de la première partie.

La syntaxe d'une règle a schématiquement la forme suivante :

focus ortho. *Chaîne phon.*
TOPH: (Cg)+ChE+(Cd){Cat}—>[ChS]
PHOT: [PhCg]+ChE+[PhCd]{Cat}—>(ChS)
focus phon. *Chaîne ortho.*

où le focus est la fenêtre d'observation décrite explicitement dans une règle.

Les contextes sont définis sur les ensembles ou directement sur les vocabulaires d'entrée et avec deux opérateurs (l'énumération ',' et la concaténation '+'). Par exemple, la règle décrivant la phonétisation [l'orthographisation] du graphème s [phonème[s]] dans le mot "tournesol" [tuʀnəsɔl] est (TOPH : R1 ; PHOT : ℜ1) :

R1: ("#"+tourne) +s+ (ol+"#") —> [s]
ℜ1: ["#"+tuʀnə] +s+ [ɔl+"#"] —> (s)

où "#"est l'ensemble des graphies marquant la frontière de mots.

Pendant la transcription, les règles seront examinées dans l'ordre où elles sont écrites dans une même classe (une classe regroupant les règles dont ChE commence par le même caractère). Ainsi la règle de phonétisation R2 : +s+ —> [s] placée après R1 étend implicitement le contexte de R2 à {V* sauf ("#"suivi de "tourne")} à gauche et {V* sauf ("ol" suivi de "#")} à droite.

Il faut souligner que les règles sont contextuelles et sont donc toujours bornées par V*. Ainsi R1 s'applique sur un sous-ensemble des chaînes définies par V*.focus (R1).V*, (i.e. V*."#"tournesol"#".V*). R2 s'applique sur un sous-ensemble des chaînes définies par V*.s.V*et plus exactement, si la grammaire est réduite à R1;R2, R2 s'applique sur les chaînes V*.{V*\("#"tourne)}.s.{V*\(ol"#")}.V*. Plus généralement une règle TOPH s'applique sur un sous-ensemble de chaînes de V*.Cg.ChE.Cd.V*, et la règle inverse PHOT s'applique sur S*.Cg.ChS.Cd.S*. Si les contextes sont vides dans la syntaxe TOPH, la règle inverse s'applique sur S*.ChS.S*. Donc, une règle donnée sans contexte dans PHOT décrit implicitement un contexte plus général, voire quelconque.

2.2. La boîte à outil TOPH

Les logiciels sont écrits en langage Pascal norme ISO transportable. Les interfaces sont développées dans l'environnement HyperTalk/HyperCard sous Macintosh. Différents outils ont dû être proposés à l'expert afin qu'il garde la maîtrise des connaissances qu'il formalise.

2.2.1. Trace d'application et statistiques

Ce premier outil fournit la trace des règles choisies et appliquées par le phonétiseur et les statistiques de d'application de toutes les règles de la grammaire, soit localement, soit en cumulant les historiques. Pour chaque règle sont spécifiés le nombre d'utilisation de cette règle dans le texte d'entrée, le pourcentage d'utilisation par rapport aux règles de la même classe, et le pourcentage d'utilisation par rapport à la totalité des règles. Si la trace permet à l'expert de détecter et d'identifier les erreurs de transcription, l'interprétation des statistiques permet ultérieurement d'associer un poids aux règles selon leur fréquence d'usage et de détecter les règles inaccessibles.

2.2.2. Trace de la grammaire

Nous avons vu que l'ordre de déclaration des règles étend implicitement la définition des contextes des règles (dépendance verticale). L'ordre de parcours de la chaîne à transcrire implique également des contraintes sur l'application des règles (dépendance horizontale). La maîtrise de ces ordres se complexifie grandement avec l'augmentation du nombre des règles. Il s'est avéré nécessaire de développer des outils pour tracer les liens entre les règles dans ces relations d'ordre et en déduire les règles redondantes et incohérentes (Ghneim 95).

Une règle est dite *redondante* s'il existe d'autres règles dans la grammaire qui traitent tout ou partie de la même chaîne graphémique (avec les contextes correspondants), en générant la même transcription phonétique, par exemple:

R3 : ("#"+sporotri) +ch+ (ose+"#") —> [k]
R4 : (tri) +ch+ (o) —> [k]

où R3 et R4 traitent la même chaîne graphique "ch" dans les deux focus "#"sporotrichose"#" et "tricho"respectivement, sachant que "tricho"⊂"sporotrichose".

Une règle est dite *incohérente* s'il existe d'autres règles, situées avant elle dans la grammaire, qui traitent tout ou partie de la même chaîne graphémique (avec les contextes correspondants), en générant d'autres transcriptions phonétiques, par exemple:

R5: ("#"+st,str,sw)+ea+—>[i], (e.g. "streamer")
R6: ("#"+st) +ea+ (k+"#") —> [ɛ]

où l'expert insère une règle qui traite un cas irrégulier R6 après une règle plus générale R5.

3. DE TOPH À PHOT

En plus des contraintes intrinsèques au formalisme TOPH, des contraintes méthodologiques ont été imposées pour l'établissement de la grammaire du français. Une grammaire TOPH pour les phénomènes de phonétisation des formes canoniques du

français, étendue ensuite aux formes fléchies, a été construite par un boostrapping sur une grammaire minimale. Elle résulte de l'exploration systématique d'un lexique représentatif du français : *le 60 000* de l'ICP associées aux entrées phonétiques référentielles du Petit Robert. Elle a été (1) établie selon des critères d'optimisation systémique (2) validée linguistiquement sur le plan diachronique. De la grammaire-lexique du français émerge le graphone, unité composite du graphème (Belrhali 96). Autour du système général de la langue, se détachent des sous-systèmes qui caractérisent des sous-lexiques de mots d'emprunts ou d'usage spécifique, résultats du transfert phonologique de langue-source vers le français (Aubergé 1996).

3.1. Contrainte de couverture

Une sous-chaîne ChE quelconque peut être réécrite par TOPH et décrite en focus sous la forme suivante:

"#".G.Cg.ChE.Cd.D."#"

où les deux sous chaînes Cg et Cd sont nécessaires pour effectuer la réécriture, G et D sont des sous chaînes de V* non explicitées, mais décrites en partie ou en tout implicitement par la relation d'ordre vertical. On voit donc que V* est limité intrinsèquement à gauche et à droite par l'étendue de l'unité linguistique maximale sur laquelle peut porter la réécriture, c'est-à-dire l'unité lexicale. En effet, la correspondance entre graphies et sons est conditionnée soit par un contexte phonotactique, soit par un contexte relatif à une unité linguistique (affixe, base et au maximum l'unité lexicale). Dans ce deuxième cas, il s'agit donc de choisir une stratégie de conception des règles :

• critère logique : le focus décrit la sous-chaîne minimale qui discrimine l'unité linguistique dans le *60 000* (e.g. dans la règle R4 : (otri) +ch+ (o) —> [k] le focus "otricho" accède au mot "sporotrichose") ;

• critère linguistique : le focus recouvre l'unité linguistique (e.g. le focus de R3).

Nous avons imposé systématiquement le critère linguistique pour l'établissement des règles TOPH (1) dans un but de lisibilité linguistique (2) dans le but conséquent de réversibilité des règles : une contrainte contextuelle logiquement suffisante dans TOPH n'est pas nécessairement suffisante dans PHOT, par contre une contrainte contextuelle linguistique est nécessaire et suffisante dans PHOT.

3.2. La phonétisation des contextes des règles de TOPH

Calculer une règle PHOT à partir d'une règle TOPH consiste à inverser la chaîne à transcrire en entrée et celle en sortie de la règle TOPH, à phonétiser les contextes gauche et droit de la règle à l'aide de la grammaire TOPH et à phonétiser les lexiques et les ensembles. Par exemple, déterminer les contextes phonétiques de la règle R7 : +oa+ (ch+"#") —> [o] dont le focus est V*.oach."#"V*, revient à déterminer la phonétisation de la chaîne "ch" dans le focus V*.oach."#".V*. Supposons que la grammaire TOPH contienne les règles R8: ("#"+coa) +ch+ ("#") —> [tʃ] et R9: +ch+ —> [ʃ] où le focus de R8 est V*.coach."#".V*, et celui de R12 est V*.V\{c}.oach."#".V*. La règle R8 n'est pas applicable pour phonétiser le contexte droit de R7 car son focus est plus spécifié que celui de R7 ; ainsi R9 sera appliquée.

Cet exemple illustre une incohérence entre critère logique et linguistique dans le choix de la couverture des règles. Or le nombre de solutions possibles associées à chaque contexte est potentiellement exponentiel : il faut appliquer toute règle dont le focus inclut le focus de V*. En choisissant systématiquement le critère linguistique, nous allons diminuer le nombre solutions de phonétisation des contextes.

Considérons les systèmes, comme VORTEX (Pecatte 92) qui mettent en œuvre des correspondances simples sons-graphies sans contexte (c'est-à-dire, ramenés à TOPH, des contextes toujours égaux à V* sur la limite de la couverture maximale, i.e. l'unité lexicale), les solutions engendrées sont multiples (facteur exponentiel) et la cohérence sur le mot est assuré par un accès lexical. Les solutions que nous engendrons par TOPH en conservant la contrainte des contextes reste, sur le plan formel, exponentiel (c'est donc le nombre de règles PHOT qui est potentiellement exponentiel) ; mais la cohérence linguistique intrinsèque au fonctionnement graphies<–>sons et véhiculé par les contraintes de méthodes d'écriture des règles résout implicitement le choix multiple.

En ce qui concerne les règles à contextes phonotactiques, par exemple la règle R10: ("Voyelle") +s+ ("Voyelle") —> [z] dont le focus est V*."Voyelle".s."Voyelle".V*, on conserve un facteur exponentiel (les solutions phonétiques des voyelles ne sont pas contraintes par ce focus). Dans le focus de la règle PHOT

ℜ10: ["Voyelle phon."] +z+ ["Voyelle p

hon"] —> [s], le nombre de sous-chaînes décrites est certes exponentiel par rapport aux sous-chaînes du focus de R10 mais est moins élevé que le nombre total de focus possibles dans V*, ce qui serait le cas pour une grammaire réduite à s—> [z]. On peut donc supposer que les solutions d'orthographisation proposées par ces règles seront plus restreintes (puisque plus contraintes) que celles proposées par une grammaire réduite à des règles du type ChE—> [ChS].

3.3. Ordonnancement de PHOT

Les règles constituant une classe de la grammaire PHOT proviennent de différentes classes dans TOPH. L'ordre vertical de TOPH ne peut pas être transposé dans PHOT. Par exemple la classe de [e] est constituée de règles provenant de la classe des graphies "a, æ, e, é, è, ê, ë, œ" :

ℜ11: ["#"+bʀ]+ɛ+[n+"#"]—>(ai) {"brain"}
ℜ12: ["#"+j] +ɛ+[n+"#"]—>(e) {"yen"}

Les règles de PHOT doivent donc être réorganisées a posteriori, contrairement aux règles TOPH qui sont ordonnées selon le choix de l'expert. Nous proposons pour cela une méthode semi-automatique analogue à celle qui permet à l'expert de détecter les redondances et les incohérences : en appliquant la procédure de trace de la grammaire PHOT une solution de classement est proposé à l'expert qui garde le choix de valider ou non cette solution.

4. CONCLUSION

Nous avons développé un environnement de « programmation » des grammaires-lexiques TOPH et PHOT, et une méthode « algorithmique » pour l'établissement des grammaires TOPH. Nous avons pu montré que la description de l'orthographisation peut être déduite de la phonétisation. Cependant l'inversion de TOPH et plus particulièrement la phonétisation par TOPH des contextes de TOPH reste un problème complexe, auquel nous avons pu apporter des solutions automatiques. Nous devons maintenant montré (1) que cette grammaire est « lisible » sur le plan linguistique (2) que les formes orthographiques obtenues sont moins nombreuses que celles obtenues avec l'inversion de TOPH sans les contextes. Les deux grammaires TOPH et PHOT seront ensuite utilisées dans un module de correction de fautes afin de comparer les deux méthodes classiques : (1) phonétiser le mot à corriger, ensuite effectuer la recherche dans un lexique par accès tolérant (distance phonétique) (2) phonétiser le mot à corriger, ensuite orthographier le résultat et enfin effectuer la recherche dans un lexique par accès tolérant (distance graphique).

5. BIBLIOGRAPHIE

Ainsworth W.A., Pell B. (1990) Connectionist Architectures for a Text-to-Speech System, *Eurospeech*, 125-128.

Aubergé, V. (1991) *La synthèse de la parole: des règles aux lexiques*, Thèse de l'université Pierre Mendès France, Grenoble2.

Aubergé V, Belrhali R. (1996) La phonétisation automatique du français : émergence de règles ou de lexiques ?, *Revue LIDIL* n° 13 du Labo. de Linguistique et Didactique des Langues étrangères Maternelles, Édition Pug.

Belrhali R. (1995) *Phonétisation automatique d'un lexique général du français : systémique et émergence linguistique*, Thèse de l'Université Stendhal, Grenoble.

Catach, N. (1984) *La Phonétisation automatique du français*, Édition CNRS, Paris .

Divay M., Guyomard M. (1977) *Conception et réalisation sur ordinateur d'un programme de transcription graphémo-phonétique du français*, Thèse de 3ème cycle, Université de Rennes.

Ghneim N., Aubergé V. (1995), Optimisation de la grammaire de phonétisation du français TOPH en vue de la correction orthographique, 3rd Int. Conf. on statistical Analysis of Textual data *JADT*, Rome.

Hunnicutt S., Meng H., Seneff S., Zue V. (1993), Reversible Letter-to-Sound Sound-to-Letter generation based on parsing word morphology, *EuroSpeech*, Germany, Vol. 2, pp.763-766.

Lacheret-Dujour A. (1990) *Contribution à l'analyse de la variabilité phonologique pour le traitement automatique de la parole continue multilocuteur*, Thèse de l'Université Paris VII.

Laporte E., Silberztein M. (1989) Vérification et correction orthographique assistées par ordinateur, *Actes de la 1ère conférence européenne sur les techniques et les applications de l'Intelligence Artificielle en milieu industriel et de service*, Hermès.

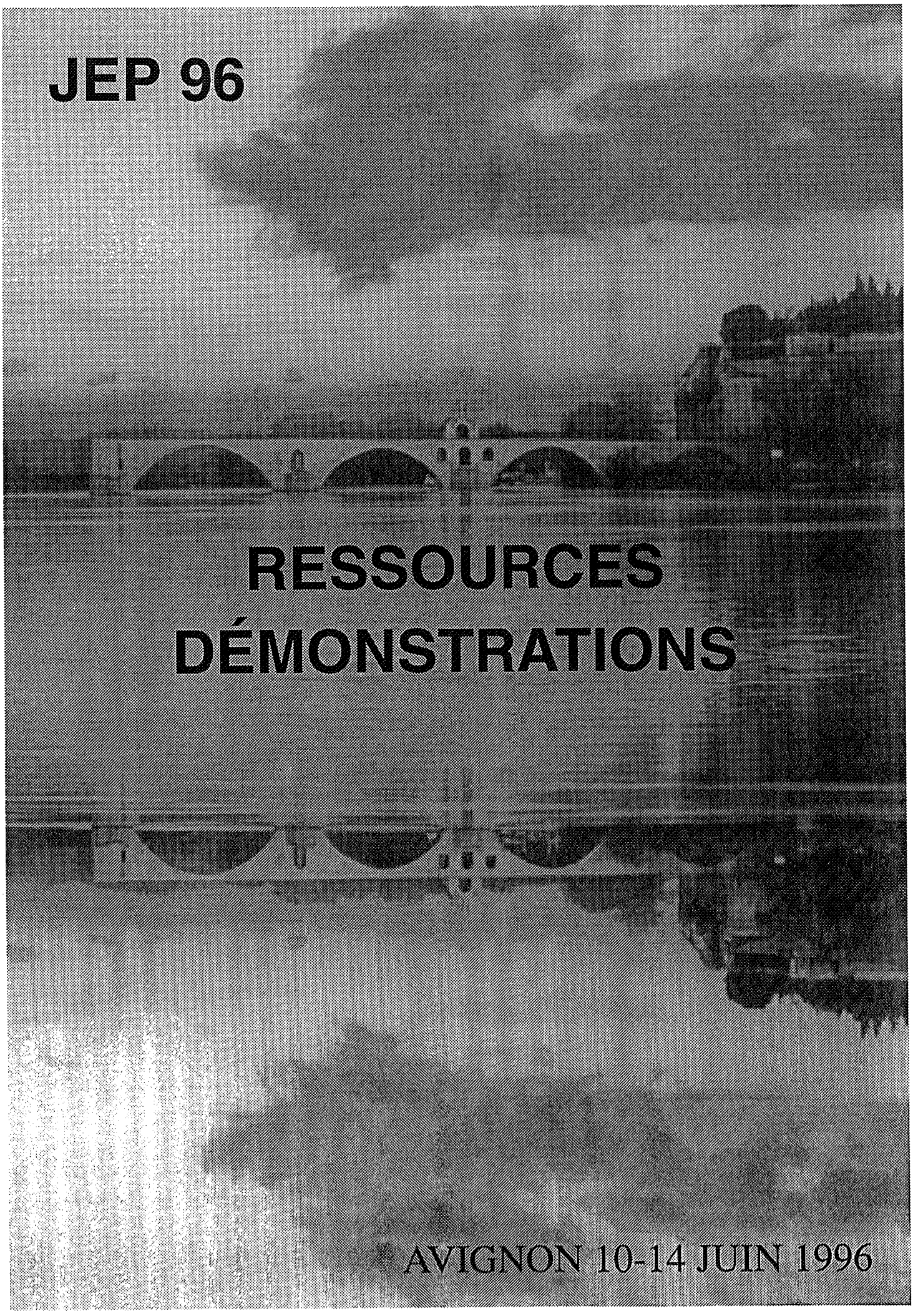
Néel F., Eskenasi M., Mariani J. (1986) Module de traduction phonétique avec variantes, *Séminaire GRECO-GALF Comm. Parlée*, Toulouse.

Pécatte J. (1992) *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, Toulouse.

Strube de Lima V.L. (1990) *Contribution à l'étude du traitement des erreurs au niveau lexico-syntaxique dans un texte écrit en français*, Thèse de l'Université Joseph Fourrier, Grenoble I.

Véronis, J. (1993). Distance entre chaînes : extension aux erreurs phono-graphiques, *Travaux de l'Institut de Phonétique d'AIX*, vol. 15, pp.217-234.

JEP 96



RESSOURCES DÉMONSTRATIONS

AVIGNON 10-14 JUIN 1996

WINPITCH: F0 EN TEMPS REEL SOUS WINDOWS 3.1

Philippe Martin
Experimental Phonetics Laboratory
University of Toronto
Carr Hall #101, St Jospeh St. Toronto, Ont. M5S 1A1
(416) 926 1300 #3279 - e-mail: pmartin@epas.utoronto.ca

ABSTRACT

The spectral comb method for pitch analysis was introduced some 15 years ago (Martin, 1981). We present here a revised version of this algorithm, allowing real time Fo analysis on a Pentium 90 MHz based computer, and integrated into a general purpose pitch analyzer windows program. This program allows for temporal, intensity and frequency editing of analyzed sentences through psola synthesis, as well as numerous tools for Fo research and teaching applications. Jitter and shimmer measurements can be made on user selectable blocks of the signal through glottal cycle marking, text can be added for model-imitation sessions in a teaching environment working in a karaoke mode, etc. The program uses a standard Sound Blaster compatible sound card as a front end.

1. INTRODUCTION

La mesure de la fréquence fondamentale du signal de parole (F0) demeure une des techniques essentielles de la recherche en phonétique expérimentale. Ses applications dans les domaines de l'enseignement de la prosodie des langues étrangères, de l' orthophonie, de la phonostylistique sont également très importantes.

La nature du signal de parole, et les conditions de bruit d'enregis-trement rendent souvent l'analyse de F0 difficile et entachée de nombreuses erreurs (Hermes, 1993). Seules les méthodes basées sur une analyse spectrale et exploitant le plus d'informations possible sur la structure harmonique du signal dans ses parties voisées présentent des performances satisfaisantes pour des locuteurs et des conditions d'enregistrement variées.

Parmi ces méthodes, la détection de Fo par peigne spectral (Martin, 1981, 1983, 1986), introduite il y a plus de 15 ans, présente par sa simplicité des possibilités intéressantes d'implanta-tion en temps réel, nécessaires pour les applications d'enseignement, et permet en recherche phonétique une utilisation performante pour le dépouillement de larges corpus d'analyse.

On présente ici une version modifiée de la méthode du peigne spectral, qui offre la particularité de fonctionner en temps réel sur des équipements standard de grande diffusion (processeur Pentium 90 MHz et une carte son Sound Blaster). Son implantation est intégrée dans un programme multifonctions pour la recherche et l'enseignement de l'intonation, fonctionnant sous Windows 3.1, et permettant, outre la mesure de Fo, de l'intensité, de la durée, etc., la synthèse du signal original après modification des paramètres de fréquence fondamentale, d'intensité et de durée, la visualisation instantanée des formes d'onde, du spectre, des marqueurs de cycle glottique, de la fonction peigne des signaux originaux et synthétisés, la présentation de courbes mélodiques en mode karaoke après adjonction de texte pour la présentation de courbes modèles dans un cadre d'enseignement, l'édition du signal par couper-coller, la mesure statistique des variations fines de F0 et de l'intensité (histogrammes du *jitter* et du *shimmer*), etc.

2. METHODE

Analyse de Fo par la méthode dite du peigne spectral

La méthode du peigne spectral pour l'estimation de Fo est basée sur la recherche d'une structure harmonique dans le spectre instantané, qui intègre à la fois les informations de fréquence et d'amplitude. Ceci permet une détection correcte de Fo même si le spectre est dépourvu de composante fondamentale. (on peut du reste noter que des méthodes spectrales comme le cepstre ne permettent pas cette détection si le signal n'a pas d'harmoniques, comme dans le cas d'un son pur).

Dans le processus d'évaluation de Fo, le spectre instantané $F(\omega)$ est tout d'abord "nettoyé" (shampouiné) en remplaçant les sommets spectraux, évalués par interpolation à partir des valeurs du spectre FFT, et remplissant certaines conditions, par des fonctions paraboliques, et le reste du spectre par des valeurs nulles. De cette manière, des composantes de bruit non liées à la structure harmonique du signal et d'amplitude faibles pourront être éliminées du calcul.

Le spectre shampouiné $F'(\omega)$ est ensuite intercorrélé avec une fonction spectrale peigne $C(\omega p,\omega)$, avec des "dents" d'amplitude décroissante An et d'intervalle fréquentiel ωp:

$$C(\omega p,\omega) = \sum An(n\omega p - \omega)$$

Le maximum de la fonction d'intercorrélation $I(\omega p)$ est atteint lorsque un grand nombre de dents du peigne coïncide sur l'axe des fréquences avec les sommets harmoniques du spectre shampouiné. Lorsque ce maximum excède un certain seuil, la valeur correspondante de l'intervalle fréquentiel est retenu comme valeur de Fo.

$$I(\omega p) = \sum An\,|F'(\omega)|$$

La nature du spectre shampouiné et le grand nombre de valeurs nulles impliquées dans le calcul fait qu'un algorithme efficace peut être obtenu en ne considérant que les valeurs non nulles dans le calcul. Ceci est obtenu aisément en ne retenant que les valeurs fréquentielle et d'amplitude de chaque sommet spectral du spectre instantané $F(\omega)$, et en remplaçant chaque sommet par une parabole appropriée dans un spectre où toutes les valeurs sont nulles au départ. La fonction d'intercorrélation $I(\omega p)$ est alors obtenue en additionnant n de ces spectres, dont les échelles sont réduites respectivement:

- sur l'axe des fréquences par un facteur n correspondant à l'ordre de la dent du peigne;

- sur l'axe des amplitudes par un facteur m correspondant à la réduction d'amplitude de la dent d'ordre n;

Le caractère voisé ou non-voisé du signal pour la fenêtre d'analyse est établi en combinant linéairement le nombre de passage par zéro et le seuil de voisement relatif à la fonction peigne $I(\omega p)$.

La courbe mélodique résultante est ensuite adoucie par un filtre médian d'ordre paramétré.

Un processeur Pentium 90 MHz réalise une analyse en 12 ms environ, utilisant une FFT de 256 points, soit une fenêtre temporelle de 23 ms à la fréquence d'échantillonnage de 11025 Hz (valeur imposée par la carte de conversion et par le standard multimédia de Windows).

Le temps réel est donc obtenu avec un facteur de recouvrement de fenêtres d'analyse d'environ 2. Lorsque le nombre de pixels de l'écran entre deux valeurs successives est supérieur à 1 (cas d'une grande résolution associée à une durée totale d'écran faible, par exemple 980 pixels pour une durée d'une seconde), les valeurs intermédiaires inscrites sont obtenues par interpolation linéaire.

Le programme permet également un mode d'analyse avec un facteur de recouvrement variable, permettant l'examen détaillé de l'évolution des paramètres prosodiques ainsi que la visualisation du spectre FFT. L'obtention du temps réel dépend alors des caractéristiques du processeur utilisé.

Nouveau modèle du peigne

Dans l'implantation du programme sous Windows, une fenêtre permet de visualiser instantanément la forme d'onde, le spectre, la fonction peigne, etc. en n'importe quel point du signal. Les causes éventuelles des erreurs d'estimation de la fréquence fondamentale peuvent ainsi être rapidement et facilement évaluées.

L'analyse des erreurs pour des types de voix variées et dans des conditions d'enregistrement non optimales ont conduit à l'examen systématique des effets de la modification de certains paramètres d'analyse, parmi lesquels on a retenu, pour une fréquence d'échantillonnage donnée:

- Fo minimal
- Fo maximal
- Fréquence maximale du spectre
- Seuil d'écrêtage du spectre (shampooinage)
- Nombre d'harmoniques considérés
- Nombre de dents du peigne
- Paramètre de décroissance des dents
- Seuil de voisement (sur la fonction peigne)
- Seuil de voisement par passages par zéro
- Applatissement du spectre:

De nombreuses erreurs peuvent se produire lorsque l'amplitude de la fondamentale et de la deuxième harmonique est très inférieure à celles des harmoniques supérieures (de plus de 20 dB). On a donc ajouté aux paramètres ci-dessus la possibilité d'un traitement du spectre FFT après un aplatissement spectral, dans lequel les sommets du spectre du signal reçoivent tous la même amplitude avant le calcul d'intercorrélation.

Un même ensemble de valeurs de paramètres ne donnant pas de résultats satisfaisant pour toutes les conditions d'enregistrement et tous les types de vois retenues, les différentes valeurs ont été regroupées en un ensemble de 8 vecteurs de paramètres, censés correspondre aux valeurs optimales pour l'analyse correcte des différents types de voix.

L'analyse se fait alors, pour une trame donnée, plusieurs fois consécutivement pour des ensembles de valeurs différentes du vecteur de paramètre. La valeur de Fo retenue est alors prise démocratiquement à la majorité simple des valeurs mesurées.

3. REFERENCES

Hermes, D. (1993) "Pitch Analysis", in *Visual Representations of Speech Signal*, M. Cooke, S. Beet and M. Crawford, ed. Wiley & Sons, Chichester, pp. 3-26.

Martin, Ph. (1981) "Mesure de la fréquence fondamentale par intercorrélation avec une fonction peigne", *Actes des XIIèmes JEP*, Montréal.

Martin, Ph. (1983) "Real Time Fundamental Frequency Analysis using the Spectral Comb Method", *Proceedings of the Xth Congress of Phonetic Sciences*, Foris.

Martin, Ph. (1986) "A Fast Spectral Comb Algorithm for Fo Detection", *Proceedings from the 12th International Congress of Acoustics, Toronto*, August 1986, paper A6-9.

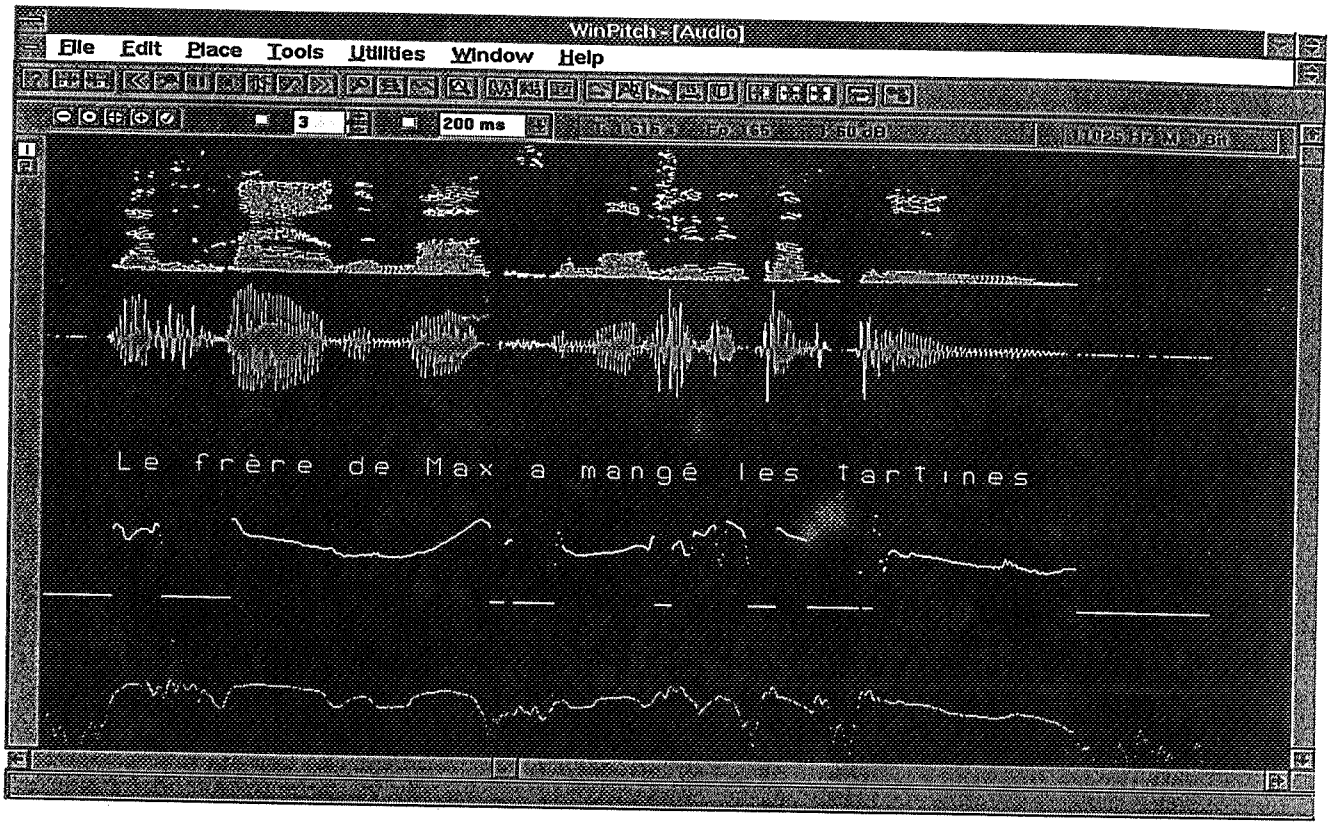


Fig.1:écran principal de WinPitch, avec spectrogramme, courbes de signal, de F0 et d'intensité, et texte.

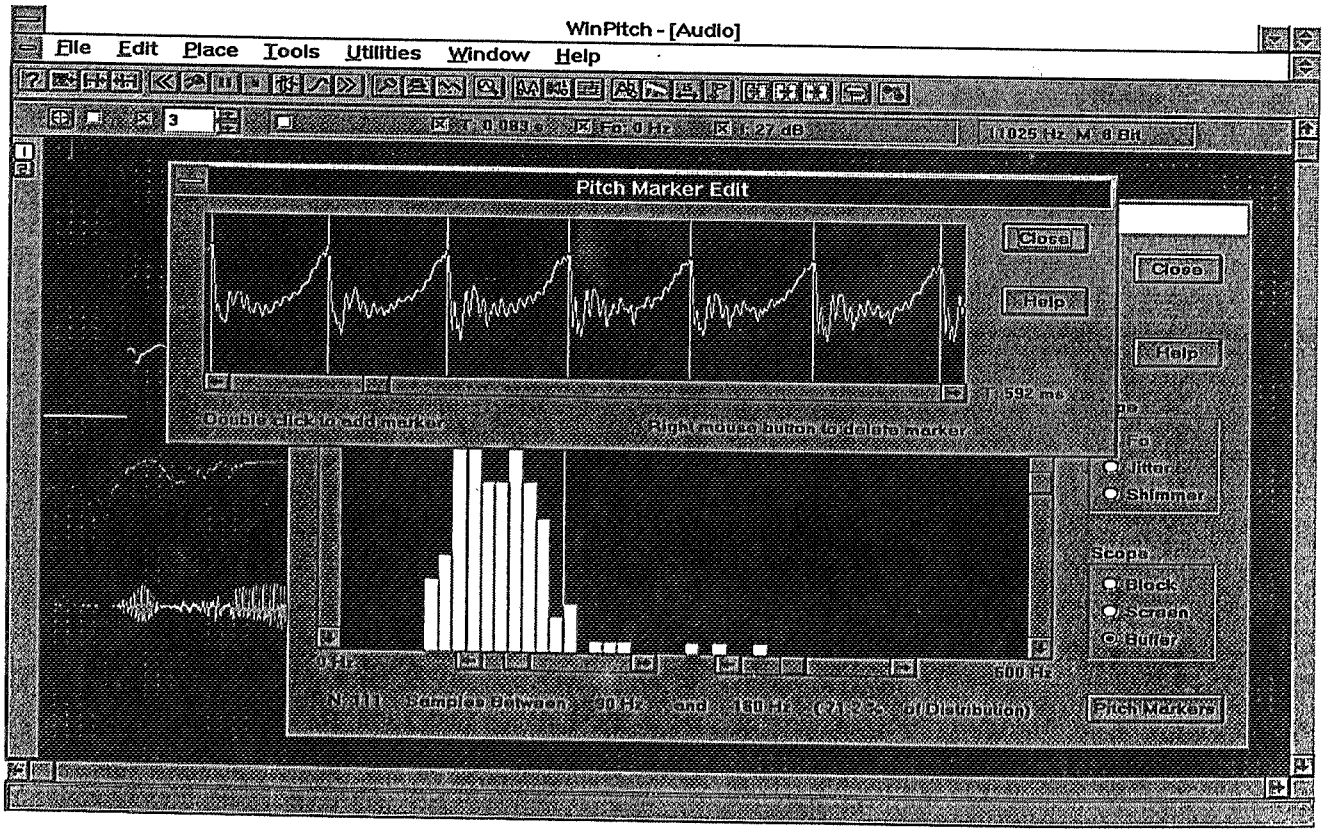


Fig. 2.: Fenêtres d'édition des marqueurs de pitch et de statistiques des variations fines de F0.

Le projet MBROLA : Vers un ensemble de synthétiseurs vocaux disponibles gratuitement pour utilisation non-commerciale

T. DUTOIT (+), V. PAGEL (*)

(+) Faculté Polytechnique de Mons, TCTS Lab, 31, Boulevard DOLEZ, B-7000 Mons, Belgique.
Tél : /32/65/374133 - Fax : /32/65/374126 - e-mail : mbrola@tcts.fpms.ac.be
(+) CRIN-CNRS & INRIA-Lorraine, BP 239, 54506, Vandoeuvre-lès-Nancy, France.

ABSTRACT

The aim of the MBROLA project, recently initiated by the TCTS Lab of the Faculté Polytechnique de Mons (Belgium), is to obtain a set of speech synthesizers for as many voices, languages and dialects as possible, free of use for non-commercial applications. The ultimate goal is to boost up academic research on speech synthesis, and particularly on prosody generation, known as one of the biggest challenges taken up by Text-To-Speech synthesizers for the years to come.

Central to the MBROLA project is MBROLA v2.00, a speech synthesizer based on the concatenation of diphones. Executable files of this synthesizer have been made freely available for many computers/operating systems.

We describe here the terms of participation to the project, as a user, as an associated developer, or as a database provider.

1. INTRODUCTION

Le laboratoire TCTS de la Faculté Polytechnique de Mons vient de lancer un projet ambitieux dans le domaine de la synthèse de la parole. Son but est de réunir un ensemble de synthétiseurs de parole pour le plus grand nombre de langues, de dialectes, et de voix possibles, et les fournir gratuitement à la communauté scientifique, pour toute utilisation non-commerciale et non-militaire. Le but ultime de ce projet est d'amplifier la recherche académique en synthèse de parole, et plus particulièrement en génération de prosodie, qui reste un problème majeur pour les années à venir.

Au coeur de ce projet : MBROLA v.2.00, un synthétiseur de parole basé sur la concaténation de diphones, dont des exécutables sont maintenant disponibles gratuitement pour de nombreuses machines. MBROLA 2.00 prend en entrée une liste de phonèmes associés à des informations prosodiques (durée des phonèmes et représentation linéaire par morceaux de la courbe de pitch : on ne pré-suppose aucun modèle linguistique) et produit des échantillons de parole sur 16 bits (linéaires), à la fréquence d'échantillonnage de la base de données de diphones utilisée. (Il ne s'agit donc PAS d'un synthétiseur à partir du texte (TTS), puisqu'il n'accepte pas de texte en entrée.)

Le logiciel est fourni avec une première base de données de diphones (FR1 : une voix masculine française) et des fichiers phonétiques qui en démontrent la qualité. Nous nous attendons à ce que le nombre de langues/voix disponibles à travers ce projet croisse rapidement, en conséquence de la politique de mise en commun de bases de données de diphones que nous avons cherché a favoriser dans ce projet.

Nous décrivons brièvement les paragraphes qui suivent les avantages de l'algorithme MBROLA, comment on peut copier le programme MBROLA V2.00, comment on peut l'utiliser et quelles sont ses entrées-sorties, ainsi que ses limites. Nous décrivons pour terminer les modalités de participation à ce projet, en tant qu'utilisateur ou en tant que fournisseur de corpus.

2. L'ALGORITHME MBROLA.

Le programme MBROLA v.2.00 utilise une technique de synthèse connue elle-même sous le nom de MBROLA[1] (Multi-Band Resynthesis OveLap Add), non encore publiée,

[1] Une demande de brevet est en cours, pour le compte de la Faculté Polytechnique de Mons.

et s'inspirant de l'algorithme MBR-PSOLA (Dutoit & Leich, 1993 ou, avec plus de détails, Dutoit, 1996). Cet algorithme produit de la parole de synthèse par concaténation de diphones (ou de triphones, de polyphones, etc...). Comme TD-PSOLA[2] (Time-Domain Pitch-Synchronous OveLap Add, Moulines & Charpentier, 1989), il réalise cette opération directement dans le domaine temporel. Au contraire de TD-PSOLA, cependant, il utilise pour ce faire une base de données spécialement adaptée aux besoins du synthétiseur, et obtenue après une analyse/re-synthèse assez complexe basée sur le modèle hybride Harmonique/Stochastique (Griffin & Lim, 1988; Abrantes *et al.*, 1991). Ceci permet à MBROLA de conserver la simplicité de calcul des méthodes de synthèse dans le domaine temporel tout en lui conférant la souplesse des méthodes paramétriques. Il résulte que cette technique présente les avantages suivants :

- Sa charge de calcul est maintenue à un niveau très bas (7 opérations par échantillon en moyenne), tout en permettant malgré tout au synthétiseur de lisser les discontinuités spectrales apparaissant de part et d'autre des points de concaténation. Ceci conduit, pour une fréquence d'échantillonnage de 16 kHz, à une synthèse en temps réel sur processeur Intel486
- La parole de synthèse obtenue est par conséquent très "fluide" (voir les comparaisons avec d'autres techniques de synthèse dans Dutoit & Leich, 1994), de sorte que même une base de diphones (par opposition, par exemple, à des polyphones) permet d'obtenir une excellente qualité de synthèse sans avoir recours à une procédure d'essais-erreurs longue et fastidieuse durant laquelle les "mauvais" diphones (ceux qui introduisent les plus grandes discontinuités spectrales) seraient éliminés et remplacés par de nouveaux enregistrements. La qualité disponible avec la base de données FR1, par exemple, est celle obtenue directement après enregistrement de la base de diphones et analyse/re-synthèse (automatique). Cette caractéristique importante est précisément celle qui fait de MBROLA un candidat idéal pour le rassemblement de nombreuses bases de données de diphones existantes ou à créer, ce qui constitue précisément le but de ce projet.
- Enfin, il est possible de coder les bases de données MBROLA de façon très efficiente (moins de 50 kbps pour une fréquence d'échantillonnage de 16 kHz, à comparer aux 256 kbps du codage linéaire sur 16 bits) et avec un sur-coût de calcul à la synthèse très minime.

3. LE PROGRAMME MBROLA V2.00

MBROLA v2.00 reçoit en entrée un fichier (ou un pipe) fournissant la suite des sons à synthétiser et la prosodie qu'on veut leur conférer. Le fichier (ou pipe) de sortie comprend des échantillons sur 16 bits linéaires à la fréquence d'échantillonnage de la base de données utilisée (16 kHz pour FR1). Le fichier audio généré répond au format de son extension .wav; .au, .aiff, ou .raw. Le format du fichier d'entrée est très simple. Le fichier `bonjour.pho`, fourni avec d'autres exemples, contient simplement :

```
_ 51 25 114
b 62
on 127 48 170
j 110 53 116
ou 211
r 150 50 91
_ 91
```

Chaque ligne contient le nom d'un phonème (ou d'un allophone), une durée, et éventuellement une suite de points de description du pitch, composées de deux valeurs : la position du point relativement à la durée du phonème (en %) et la fréquence qui y est associée. Ainsi, la première ligne de `bonjour.pho` indique au synthétiseur de produire un silence de 51 ms et de mettre un premier point de pitch valant 114 Hz à 25 % des 51 ms. Les points de pitch permettent de définir une courbe d'intonation linéaire par morceaux. Cette courbe peut être continue, vu que le programme décide lui-même de ne pas tenir compte du pitch lors de la synthèse d'un segment non-voisé.

[2] PSOLA-TD® est une marque déposée par France Télécom, qui a également déposé un brevet dans de nombreux pays sur le procédé qui y est relatif.

MBROLA v.2.00 n'est pas une "version de démonstration" qui restreindrait volontairement l'utilisation du programme. Ses seules limitations (actuelles) sont les suivantes :

1. Le pitch est limité à une plage de deux octaves; cette limitation sera levée dans des version ultérieures.

2. Bien que trois points de pitch par phonème suffisent à produire une parole de très bonne qualité, le programme en accepte jusqu'à 20, ce qui permet son utilisation en parole chantée (notamment pour la reproduction des vibratos).

3. Les phonèmes peuvent être synthétisés avec une durée maximale qui dépend de la fréquence fondamentale utilisée pour la synthèse. Plus la fréquence est élevée, plus la durée maximale est faible. Dans le cas de FR1, pour une fréquence de synthèse de 133 Hz, la durée maximale est de 7.5 sec; elles passe à 15 sec pour 66.5 Hz et à 3.75 sec pour 266 Hz.

4.Enfin, bien que les points de pitch soient facultatifs, le synthétiseur refusera de produire des suites de plus de 250 phonèmes sans information prosodique.

Des versions exécutables sont disponibles à partir de sites miroir en France, Suisse, et US pour les machines/OS suivantes :

SUN/SunOS5.4 (Solaris2.4)
SUN4 (SunOS4.xx)
HPUX9.0 and HPUX10.0
VAX/VMS V6.2
DECALPHA(AXP)/VMS 6.2
PC/DOS6
PC/WIN31
PC/LINUX, format a.out
PC/Solaris2.4
Next/NextStep

Le programme est par ailleurs fourni avec quelques utilitaires qui permettent d'en écouter les résultats sur différentes machines, en utilisant, lorsque c'est possible, les "pipes". Sur SUN, par exemple, une commande du type :

```
mbrola fr1 bonjour.pho -.au | audioplay
```

permet de lancer la synthèse du mot bonjour dont les informations phonétiques sont fournies par le fichier `bonjour.pho`, avec la base de données FR1, et d'en envoyer instantannément le résultat sur le haut parleur de la SUN.

4. PARTICIPATIONS AU PROJET.

Les participations au projet MBROLA peuvent être de trois types : en tant qu'utilisateur, que développeur associé, ou en tant que fournisseur de base de données.

4.1. Utilisateurs.

Le programme MBROLA v2.00 peut être utilisé librement pour toute application non-commerciale et non militaire (suivant les termes d'un accord de licence accompagnant le programme). Une mailing liste a été mise à la disposition des utilisateurs :

`mbrola-interest@tcts.fpms.ac.be`

Elle permet aux auteurs du logiciel de diffuser diverses informations sur l'état d'avancement du projet, la mise à disposition de nouvelles bases de données, etc, ainsi qu'un échange de points de vue entre utilisateurs sur le logiciel et la mise en commun de fichiers phonétiques (extension .pho).

4.2. Développeurs associés.

La licence accompagnant le logiciel MBROLA v2.00 permet également le développement d'applications faisant usage du logiciel et de ses bases de données, dans le cadre strict du projet MBROLA : les applications résultantes doivent être mises gratuitement à disposition sur internet et annoncées via la mailing liste `mbrola-interest` et toute commercialisation est soumise à un accord préalable avec les auteurs du logiciel MBROLA.

Ceci devrait permettre à bon nombre d'utilitaires vocaux gratuits de faire leur apparition dans le cadre de ce projet, et en particulier dans le secteur de l'aide aux handicapés.

4.3. Fournisseurs de corpus.

Un des intérêts majeurs du projet MBROLA, et sans aucun doute son aspect le plus original, tient à son ambition à vouloir couvrir un nombre croissant de langues et de voix. Le projet lui-même a été organisé de

façon à inciter les laboratoires de recherche à mettre leurs bases de données de diphones à disposition, ou à en créer dans le cadre du projet. Les termes de cet arrangement peuvent être résumé de la façon suivante :

1. Nous nous engageons à n'utiliser les bases de données de diphones que pour les adapter, gratuitement, au format MBROLA et à détruire les originaux.

2. Nous intégrons la nouvelle base de données au projet MBROLA pour utilisation non-commerciale et non-militaire. En retour, nous cédons gratuitement et exclusivement nos droits sur l'utilisation commerciale de la base de données au laboratoire qui a prêté ses données, à la condition que cette base de données soit exclusivement utilisée avec le logiciel MBROLA.

L'algorithme MBROLA se prête par ailleurs fort bien à ce genre d'opération. Comme on l'a déjà signalé, en effet, l'opération de re-synthèse MBE, sur laquelle tout l'algorithme de synthèse est bâti, est entièrement automatique. De plus, les possibilités de lissage spectral performant offertes par cet algorithme permettent d'obtenir rapidement une excellente qualité de synthèse, ce qui permet de minimiser le coût afférent à la création de nouvelles bases de données en vue de leur adaptation au format MBROLA.

5. CONCLUSIONS

Par ce projet, nous espérons créer un rassemblement des scientifiques utilisateurs et développeurs de synthèse vocale autour du synthétiseur MBROLA. Ceci devrait permettre une plus grande diffusion des outils de test indispensables à la mise au point d'outils de phonétisation automatique et de génération de la prosodie, et par là même un accroissement de ce type de recherches, dont on sait qu'elles conditionnent grandement l'avenir de la synthèse vocale.

Comme le faisaient en effet récemment remarquer Cole *et al.* dans un rapport sur l'état de la recherche en parole (Cole *et al.*, 1995) : *"Les résultats d'une subsidiation accrue de la recherche en parole par l'industrie se font cruellement sentir. Au contraire de ce qui se passait il y a une dizaine d'années, il est aujourd'hui difficile d'avoir accès à un système de synthèse performant et permettant un contrôle total de ses paramètres".* Nous mettrons toute notre énergie à changer cet état des choses.

Pour plus de détail, consulter la page MBROLA sur WWW :

`http://tcts.fpms.ac.be/synthesis/mbrola.html`

6. REFERENCES

T. DUTOIT, H. LEICH, 1993, "MBR-PSOLA : Text-To-Speech Synthesis based on an MBE Re-Synthesis of the Segments Database", *Speech Communication*, Elsevier Publisher.

T. DUTOIT, 1996, *An Introduction to Text-To-Speech Synthesis,* forthcoming textbook, Kluwer Academic Publishers.

E. MOULINES, F. CHARPENTIER, 1989, "Pitch Synchronous waveform Processing techniques for Text-To-Speech Synthesis using diphones", *Speech Communication*, Vol. 9, n°5-6.

D.W. GRIFFIN, J.S. LIM, 1988, 'Multi-Band Excitation Vocoder', IEEE Trans. on ASSP, vol. ASSP-36, pp. 1223-1235.

A.J. ABRANTES, J.S. MARQUES, I.M. TRANSCOSO, 1991, "Hybrid Sinusoïdal Modeling of Speech without Voicing Decision", *EUROSPEECH 91*, pp. 231-234.

T. DUTOIT, H. LEICH, 1994, "Synthèse de la parole de haute qualité à partir d'un texte : une comparaison de quatre algorithmes candidats", *actes des XXèmes Journées d'Etudes de la Parole*, pp. 69-74.

R. COLE, L. HIRSCHMAN, L. ATLAS, M. BECKMAN, A. BIERMANN, M. BUSH, M. CLEMENTS, J. COHEN, O. GARCIA, B. HANSON, H. HERMENSKY, S. LEVINSON, K. McKEOWN, N. MORGAN, D.G. NOVICK. M. OSTENDORF, S. OVIATT, P. PRICE, H. SILVERMAN, J. SPITZ, A. WAIBEL, C. WEINSTEIN, S. ZAHORIAN, V. ZUE, 1995, "The challenge of spoken language systems : research directions for the nineties", *IEEE Trans. on Speech and Audio Processing*, vol. 3, n°1, pp. 1-21.

LE SYSTEME PHYSIOLOGIA

Bernard TESTON
Laboratoire Parole et Langage Université de Provence Aix en Provence
e-mail teston@lpl.univ-aix.fr

ABSTRACT

PHYSIOLOGIA is a PC-compatible computer-driven work station designed to record, edit, and process acoustic speech signals in relation to the corresponding physiological signals. These signals are obtained from various sources (flow-rate transducers, pressure transducers, position and movement gauges, electrodes, microphones, and laryngophones, etc.) which vary in bandwidth.

The PHYSIOLOGIA system was developed by the Institute of Phonetics in Aix-en-Provence, France (Laboratoire "Parole et Langage", URA 261 CNRS), as a tool for research on the physiological mechanisms of speech production.. It consists of an acquisition system equipped with various transducers and the signal editing and processing software PHONEDIT.

Introduction

PHYSIOLOGIA est une station de travail sur micro-ordinateur de type PC, conçue spécialement pour enregistrer, éditer, et traiter des signaux acoustiques de parole en relation avec des signaux physiologiques. Ces signaux sont de toutes origines (capteurs de débit, de pression, de positionnement, de mouvement, électrodes, microphones, laryngophones, etc...) et caractérisés par des largeurs de bande de fréquence différentes.

Elle a été développée à l'Institut de Phonétique d'Aix en Provence (Laboratoire "Parole et Langage" URA 261 du CNRS), dans le cadre du contrat européen ACCOR (ESPRIT II, BRA, Action n°3279) comme outil de recherche sur les mécanismes physiologiques de production de la parole.

1. Le système d'acquisition

La station de travail est constituée par un système d'acquisition associé à différents capteurs et instruments de mesure qui constituent le matériel spécifique et par le logiciel d'édition et de traitement PHONEDIT.

Le système d'acquisition se présente sous la forme d'un interface qui peut recevoir jusqu'à 6 modules constitués, soit par une entrée acoustique, soit par un groupe de 8 entrées physiologiques, soit par un électropalatographe (EPG). La version standard de l'interface contient 2 entrées acoustiques, 16 entrées physiologiques (2 modules de 8) et un EPG. On peut à l'extrême disposer de 6 entrées acoustiques ou de 48 entrées physiologiques. Il peut être complété par un ensemble de capteurs aérodynamiques et magnétométriques (movetrack).

2. Le module d'édition

C'est le programme PHONEDIT qui fonctionne sous l'environnement de WINDOWS 3.1. ou 95 de MICROSOFT, il permet d'éditer tous les paramètres enregistrés, de les segmenter, les marquer, les mesurer et les traiter. Une grande facilité d'utilisation est obtenue grâce à l'usage d'icônes, de menus déroulants et de nombreux programmes utilitaires ainsi que la liaison directe avec les tableurs ou gestionnaires de base de données puissants sous WINDOWS. PHONEDIT peut être utilisé sans interface d'acquisition. Dans ce cas, on peut enregistrer et écouter des signaux de parole au moyen de cartes audio de type SOUND BLASTER.

PHONEDIT permet de faire; des mesures d'amplitude, des courbes intonatives, des spectres (FFT ou autres, sonagrammes), la visualisation des contacts linguo-palatins (EPG), l'interprétation de mouvements EMA (Articulograph ou Movetrack), le positionnement de marques d'étiquetage et labels variés (alphabétique, phonétique, prosodique, etc..).

MES: UN ENVIRONNEMENT DE TRAITEMENT DU SIGNAL

Robert ESPESSER

Laboratoire Parole et Langage - 29, Avenue Robert Schuman - 13621 AIX
Tél: 42 95 36 26 - Fax: 42 59 50 96 - e-mail:Robert.Espesser@lpl.univ-aix.fr

ABSTRACT

Unix/Motif environment for displaying, labelling and processing speech signal.

1. DESCRIPTION

Logiciel d'observation, d'étiquetage (MES) et de traitement du signal de parole (SIGNAIX) sous Unix.

A chaque déplacement d'une fenêtre dans le signal, ou à chaque déplacement d'un curseur sur le signal peuvent être associées en synchronie des actions diverses (« traitements liés »); par exemple: traitement du signal visble dans la fenetre ou autour du curseur, affichage de paramètres, écoute d'une portion de signal, etc...

2. CONCEPTION

C'est une boite à outils, chacun d'eux étant une commande de niveau shell, exécutable indépendamment. En général, les outils non graphiques sont de plus des « filtres » au sens unix du terme.

MES est l'outil de visualisation et d'étiquetage du signal; il gère les traitements liés, en général des scripts shell combinant des outils de base. N'importe quel ensemble de taches peut être un traitement lié, (s'il est suffisamment rapide pour ne pas nuire à l'interactivité). MES peut échanger des messages avec d'autres instances de MES, ou recevoir des messages d'autres tâches .Un traitement lié générant du signal de parole peut ainsi utiliser une autre instance de MES et lui envoyer un message de rafraichissemement de l'affichage du signal et des éventuels traitements liés de ce MES.

L'emploi de divers interpreteurs (Korn shell, TclTk, awk, etc..) rend très aisée la maintenance, la modularité et la création de nouveaux traitements liés.

L'ensemble se veut simple, très modulaire et surtout très ouvert: outils graphiques distincts des traitements, pas d'entête complexe pour les données pour rester accessible aux autres outils de base Unix.

3. PRINCIPAUX OUTILS DISPONIBLES

-graphique: affichage de données bidimensionnelles,tridimensionnelles(type sonagraphique),affichage d'electropalatogrammes, modélisation interactive du F0.

-traitement du signal: intensité RMS, densité de passage par zéro, analyse spectrale (spectre, spectre LPC, sonagramme, analyse/resynthèse par ondelettes), filtrage, detection de F0 (méthodes disponibles:AMDF, « peigne »,autocorrélation) , modélisation de F0, analyse/resynthèse (technique PSOLA), variation du débit (technique SOLA).

Environnement requis:

Unix SysV, Motif, TclTk (shell graphique);compilation et exécution testées sous Solaris 2.4 , 2.5.

logiciel disponible à:

http://www.lpl.univ-aix.fr/projects/multext/tools.html

LISTE ALPHABÉTIQUE DES AUTEURS