BIBLIOTHÈQUE
Institut de Phonétique
AIX-EN-PROVENCE

Groupement des Acousticiens de Langue Française

5èmes JOURNEES D'ETUDE

du groupe

"COMMUNICATION PARLEE"

Avec la participation de l'AFCET

VOLUME 1

Textes des Exposés

es par
RATOIRE D'INFORMATIQUE
A MECANIQUE
SCIENCES DE L'INGENIEUR
R.S.

ORSAY
15-17 Mai 1974

Groupement des Acousticiens de Langue Française

5èmes JOURNEES D'ETUDE

du groupe

"COMMUNICATION PARLEE"

Avec la participation de l'AFCET

VOLUME I

Textes des Exposés

UNIVERSITE DE PROVENCE
Bibliothèque de l'Institut de Phonétique
CENTRE D'AIX
ORSAY

Organisées par
le LABORATOIRE D'INFORMATIQUE
POUR LA MECANIQUE
ET LES SCIENCES DE L'INGENIEUR
DU C.N.R.S.

15_17 Mai 1974

Institut de Phonétique
Inventaire n° 2081
Cote n° A/SEP 5

Groupement des Acousticiens de Langue Française

5èmes JOURNEES D'ETUDE

du groupe

"COMMUNICATION PARLEE"

Avec la participation de l'AFCET

VOLUME I

Textes des Exposés

Organisées par
le LABORATOIRE D'INFORMATIQUE
POUR LA MECANIQUE
ET LES SCIENCES DE L'INGENIEUR
DU C.N.R.S.

ORSAY
15_17 Mai 1974

Les Cinquièmes Journées d'Etude du groupe "Communication Parlée" du Groupement des Acousticiens de Langue Française, avec la participation de l'Association Française de Cybernétique Economique et Technique, ont été organisées par le Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur du C.N.R.S., à ORSAY (Essonne) les 15, 16 et 17 Mai 1974.

Les thèmes de travail fixés pour cette rencontre étaient les suivants :

1 - Contraintes linguistiques
2 - Synthèse par règles
3 - Application des contraintes linguistiques à la reconnaissance automatique de la parole

Les exposés sont classés dans trois catégories : les conférences plénières, prononcées par des personnalités invitées ; les communications ordinaires, portant sur les thèmes ci-dessus ; et les communications libres, relatives à l'étude de la parole mais n'entrant pas dans les thèmes imposés. L'ensemble est suivi de deux tables rondes : l'une traite du codage et du traitement de la parole, ainsi que de la transmission à débit réduit, l'autre porte sur les modèles de perception.

Les comptes-rendus se composent de deux volumes. Le premier comprend la majeure partie des communications ordinaires et des communications libres. Le second comprend les textes des conférences plénières, les discussions et les comptes-rendus des tables rondes.

L'organisation de cette rencontre a été rendue possible par la compréhension bienveillante de Monsieur le Professeur MALAVARD, Directeur du L.I.M.S.I., et de Monsieur RENARD, Maître de Recherche, Adjoint au Directeur. Nos remerciements vont également à Monsieur MORVAN, Administrateur au C.N.R.S., et à ses services de Gif-sur-Yvette, ainsi qu'aux responsables administratifs et techniques du Campus Universitaire d'Orsay et de l'A.D.E.R.P., qui nous ont apporté une aide soutenue et efficace.

ORSAY, Avril 1974

J.S. LIENARD

Comité d'organisation

Mmes M. CHASTAGNER
F. NEEL
J. TASSOTTE

Mlles C. CHOPPY
N. MOSCA

Mrs J.S. LIENARD
J.J. MARIANI
M. MLOUKA
G. RENARD
D. TEIL

Duplication des comptes-rendus : Service de Mécanographie du C.N.R.S. - Gif-sur-Yvette

Thème n° 1

CONTRAINTES LINGUISTIQUES DE LA PAROLE

(contraintes d'ordre phonétique,
phonologique, syntaxique, sémantique)

Pierre JUBAN
Laboratoire Interdisciplinaire de Recherches Linguistiques
U.E.R. du Langage
Université de Haute Bretagne
RENNES

C.E.I./C.S.I.
C.N.E.T. LANNION

Relation entre Phonologie et Phonétique

Résumé

Le but de cette communication est d'essayer de définir les rapports entre phonétique et phonologie, en particulier de faire état d'une approche originale faite dans le cadre du Laboratoire Interdisciplinaire de Recherches Linguistiques de l'Université de Haute Bretagne à partir des thèses théoriques avancées par le Professeur Jean GAGNEPAIN, et qui s'oppose fondamentalment aux thèses génératives qui dominent en ce moment.

Abstract

The aim of this paper is to try to give a definition of the relationship between phonetics and phonology. It states an original approach to this problem, developed by the Laboratoire Interdisciplinaire de Recherches Linguistiques (L.I.R.L.) of the Université de Haute Bretagne, based on the theoretical theses which have been forwarded by Professor Jean GAGNEPAIN. They are basically opposing the generative approach which currently dominates linguistic studies.

Les thèses que nous présentons ici sont développées à partir des travaux du professeur Jean GAGNEPAIN, linguiste, et du professeur Olivier SABOURAUD, neuro-psychiatre, qui ont fourni les bases théoriques de ce qui suit.

Si la distinctior phonologique/phonétique semble presque partout admise, y compris par l'ingénieur, la manière de la définir et les conséquences pratiques que l'on en tire sont beaucoup moins claires.

Il semble donc qu'avant de définir les contraintes résultant de la phonétique et de la phonologie, il soit bon d'en définir les rapports et d'en inscrire les domaines respectifs.

Ce que familièrement on appelle "son" en se référant au langage naturel, relève pour nous d'un processus analytique double. On constate que dans l'utilisation du langage, il y a ce que la langue dit en nous par une analyse implicite du non-encore dit, et d'autre part, ce que nous disons avec cette langue en organisant ce que nous avons à dire par la réinsertion explicite d'un modèle pré-existant dans la matérialité, phonique pour ce qui concerne notre propos, que suppose toute parole prononcée.

Pour nous donc, tout acte de parole suppose ces deux processus analytiques que nous appelons d'une part "grammaire" pour le processus implicite (ce que la langue dit en nous), et d'autre part "rhétorique" (ce que nous disons par la langue, ou réinvestissement dans la matérialité).

La définition de ces deux processus analytiques ne relève pas de la spéculation métaphysique, mais bien au contraire elle trouve sa justification dans l'étude des faits pathologiques de langage, et notamment de l'aphasie. L'aphasie se définit pour

nous par une atteinte de la capacité d'analyse implicite, ou grammaticale, atteinte qui déplace et réduit le jeu réciproque, dialectique de ce double processus analytique, et ainsi nous le rend "palpable".

En particulier, et sans rentrer dans les détails de l'observation clinique, on est amené à définir une aphasie phonologique, qui prend deux aspects; d'une part, l'aphasie dite Wernicke qui respecte l'enchaînement en phonèmes mais qui se caractérise par un trouble portant sur le choix des traits pertinents, et d'autre part l'aphasie dite de Broca, qui se caractérise par un trouble portant sur l'enchaînement en phonèmes sans que la capacité de choix des traits pertinents soit atteinte.

C'est donc l'observation clinique qui nous permet de définir les domaines de la phonologie et de la phonétique non pas en termes de substance sonore mais en termes de processus analytiques, l'aphasie étant un trouble spécifique du processus analytique implicite que nous appelons grammaire.

Cette grammaire, qui fait du langage un fait spécifiquement humain, est une formalisation inhérente au langage même; formalisation qui est totalement indépendante de la formalisation explicite du linguiste. Pour nous donc, la phonologie est l'étude de la grammaire du signifiant, de ce qui analyse implicitement en nous le "son" pour en faire du signifiant.

On sait qu'il est impossible de découper et nommer les "sons" d'une langue à partir de leurs qualités physiques puisque en tant que tel le continu sonore est indifférenciable. En baptisant les "sons", nous ôtons non seulement ce que le continu sonore a de particulier et d'appréhendable dans une réalisation particulière de ces "sons", mais surtout nous introduisons ces sons dans un système de relations qui les limitent et les opposent,

système qui ne correspond pas à l'organisation de leur réalité physique. Cependant, le processus inverse de réinsertion explicite de ce système de relations, dans la matérialité de l'acte de parole, tend vers la correspondance terme à terme, sans jamais totalement y parvenir, entre le signifiant et le "son". Pour nous la phonétique est l'étude de ce processus de réinvestissement dans la matérialité.

Au point de vue de la phonologie, ou étude de la forme du son, la définition du signifiant consistera à déterminer un réseau d'inter-relation dans lequel on définira chaque élément de manière négative, soit par opposition (traits pertinent) soit par contraste (phonèmes, tout aussi pertinents), aux autres éléments du système; et ce, en fonction du critère de pertinence. Ce critère de l'analyse du signifiant est à chercher dans l'autre face du signe, dans le signifié (dont l'étude constitue pour nous la sémiologie). Cela ne veut pas dire qu'il y ait subordination du phonologique au sémiologique: en effet, si l'on distingue deux éléments du signifiant c'est qu'ils permettent de distinguer deux signifiés, et l'on ne saurait distinguer deux signifiés sans qu'ils soient marqués par une différence de signifiant, à moins que l'on se réfère à une métaphysique du sens ou de manière plus moderne à une théorie de universaux qui nous semble à tous égards insoutenable.

Nous définissons l'analyse phonologique, elle aussi, comme double, à la fois taxinomique et générative. Le processus analytique taxinomique définit les identités oppositionnelles que sont les traits; son altération pathologique est la cause de ce que nous appelons aphasie de Wernicke. Le processus analytique génératif définit les unités contrastives ou phonèmes; son altération pathologique est la cause de l'aphasie de Broca.

Seule l'intégrité du double processus taxinomique et

génératif permet une analyse correcte du "son" en signifiant. Nous pensons donc que R. Jakobson n'a pas raison de définir le processus taxinomique comme analyse et le processus génératif comme synthèse; l'analyse pour nous est double: taxinomique et générative, corrélativement.

Au point de vue phonétique, qui est donc pour nous l'étude du réinvestissement rhétorique, à l'axe taxinomique de la phonologie nous faisons correspondre la prononciation, qui est l'étude de la réalisation matérielle et de la variation sonore des traits, et à l'axe génératif la syllabation qui détermine la "valeur" que prennent les phonèmes (en particulier la discrimination entre consonnes et voyelles, qui n'est pas à notre sens phonologique). La phonétique est donc l'étude de la prononciation et de la syllabation.

En conclusion ...provisoire:

La phonologie est l'étude de la structure implicite qui sous-tend le signifiant dans son double processus taxinomique et génératif; les définitions qu'elle donne sont énoncées en terme d'opposition et de contraste, c'est-à-dire négativement. Cette approche de la phonologie n'a rien donc de commun avec celle qui est définie dans le cadre de la phonologie générative.

La phonétique est l'étude de la structuration positive dans le double processus de phonation et de syllabation; l'étude de la phonétique dans cette optique reste à faire, bien que certains travaux déjà effectués puissent y prendre place mais une fois qu'ils seront définis et redistribués dans le cadre théorique. Cette approche de la phonétique n'a rien non plus de commun avec celle qui est définie dans le cadre de la phonologie générative, qui la réduit à une étude de la représentation finale de la

composante phonologique.

Toutefois, phonologie et phonétique n'épuisent pas l'étude des productions sonores lié au langage, et toute une partie, que nous désignons par le terme de "phonique" (qui est du ressort du physicien pour l'aspect acoustique et du physiologue pour l'aspect articulatoire) doit prendre sa place pour compléter l'ensemble. Ce dont il s'agit en fait, c'est de bie définir les tâches respectives du phonologue, du phonéticien, du physicien et du physiologue, non en fonction des catégories socio-professionnelles qu'ils incarnent, mais en fonction d'une approche théorique qui rende compte des phénomènes et qui soit expérimentalement vérifiable (par la clinique par exemple) à tous les niveaux de l'étude des productions sonores liés au langage.

INFLUENCE DU CONTEXTE VOCALIQUE SUR LA PERCEPTION DU VOISEMENT DES OCCLUSIVES .

W. SERNICLAES
INSTITUT DE PHONETIQUE
UNIVERSITE LIBRE DE BRUXELLES

P. BEJSTER
INSTITUT DE LOGOPEDIE
G H L I N

Résumé.

L'influence de la durée de la tenue sur la perception du voisement des occlusives intervocaliques a été étudiée dans 4 contextes vocaliques symétriques : (/a/, /ɛ/, /e/ et /i/). Les résultats d'une expérience d'identification montrent que le poids perceptif de la durée de la tenue décroît en fonction du degré de fermeture des voyelles adjacentes.
La dégradation du pouvoir distinctif de la durée de la tenue semble être compensée par l'intervention d'un autre indice - le délai d'établissement du voisement (VOT).

Summary.

This experiment is aimed at defining the perceptual weight of one of the acoustic correlates of voicing in French plosives - the duration of the closure which is longer for unvoiced consonants. Four sets of symmetrical VCV sequences are investigated, the vowels bounding the consonant being /a/, /ɛ/, /e/ or /i/.
Results show that the importance of the closure duration depends on the vocalic environment, diminishing as the closedness of the vowel increases.
When the stop is bounded by closed vowels the VOT is more distinctive at the acoustical level, and is probably an important cue for the perception of voicing in these contexts.

I. INTRODUCTION.

On sait que la contribution d'un indice acoustique à la perception d'un trait phonétique peut dépendre du contexte dans lequel il s'insère.

L'existence de "biais contextuels" (FANT, 1967) est le reflet des phénomènes de coarticulation qui ont été mis en évidence au niveau de la production (OHMAN, 1966).

L'objet de ce travail est d'évaluer la prégnance perceptive de l'un des corrélats acoustiques de l'opposition voisé/ non-voisé de l'occlusive intervocalique - la durée de la tenue, plus brève pour les voisées - dans quatre environnements vocaliques symétriques : /a/, /ɛ/, /e/ et /i/.

Des investigations antérieures (WAJSKOP et SWEERTS, 1973; SERNICLAES, 1973) ont montré l'importance de ce paramètre temporel pour des segments /a/ + occlusive extraits de séquences VCV.

Ces travaux ont également attiré l'attention sur l'intervention possible d'autres indices, et notamment de ceux situés après l'explosion. Aussi l'influence du contexte vocalique sera-t-elle envisagée dans le cadre de séquences VCV complètes dont on va modifier la durée de la tenue tout en conservant intact l'ensemble des autres éléments acoustiques.

II. PROCEDURE

Un traitement sur ordinateur nous a permis d'exciser des portions croissantes de la tenue de séquences V + consonne sourde + V sans affecter les autres indices.

Les réductions temporelles de la tenue ont été normalisées en considérant le

$$\text{rapport} \frac{\text{durée tenue}}{\text{durée voyelle précédente}}.$$

Pour chacune des 12 séquences VCV (4 voyelles x 3 consonnes) six réductions successives ont été opérées, correspondant à des rapports temporels $\frac{\text{tenue}}{\text{première voyelle}}$ qui s'échelonnent régulièrement entre .90 et .10.

Quatre séries expérimentales ont été construites, une pour chacun des 4 contextes /a/, /ɛ/, /e/; et /i/.

On a demandé à 19 sujets francophones d'identifier ces consonnes en ventilant leur choix sur l'une des six réponses /p/, /t/, /k/, /b/, /d/, /g/.

Les séries expérimentales comprenaient également des séquences V + $C_{voisé}$ + V et étaient présentées en 4 séances différentes réparties en 2 mois.

III. RESULTATS ET DISCUSSION.

L'ensemble des résultats est présenté dans la figure 1. Ils se caractérisent essentiellement par un écart très important entre les taux de réponses correctes ("sourdes") obtenus pour le contexte /a/ et ceux obtenus pour les contextes vocaliques mi-fermés et fermés.

Pour les séquences /a/ + occlusive + /a/ la réduction temporelle de la tenue entraîne une décroissance considérable du taux de réponses sourdes. Dans 2 cas sur 3 (/apa/ et /ata/) le score est proche du minimum absolu de 0 % pour la tenue la plus brève.

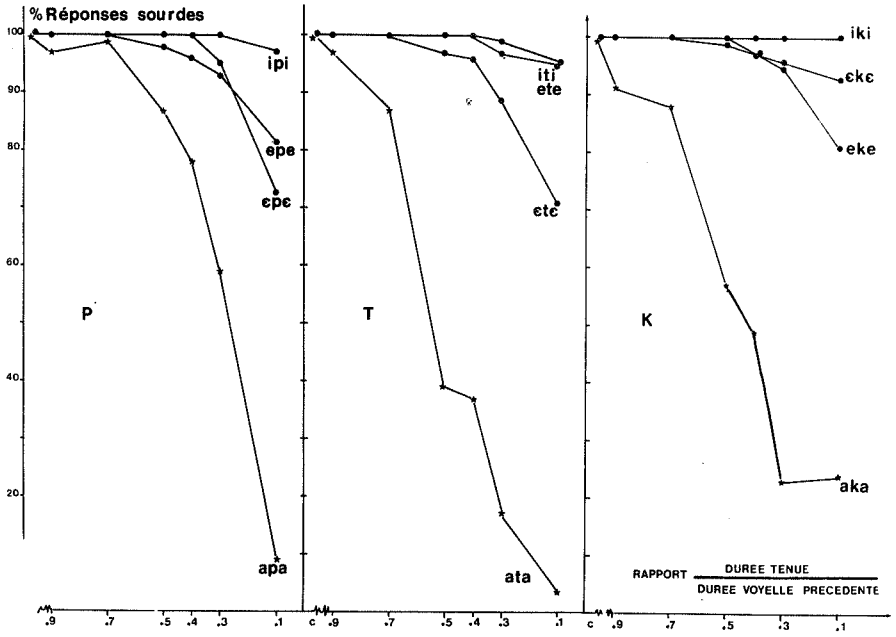


Figure 1 : Résultats de l'expérience d'identification.
En abcisse : rapport temporel tenue consonantique/voyelle précédente.
En ordonnée : pourcentage de réponses sourdes obtenu pour chaque stimulus.

Dans le contexte /i/ l'influence de la durée de la tenue est pratiquement inexistante. Pour chacune des séquences /ipi/, /iti/ et /iki/ le score reste dans le voisinage de 100 % et ce même pour les durées les plus brèves.

Pour les deux autres contextes (/ɛ/ et /e/), on observe dans certains cas une légère décroissance du pourcentage de réponses sourdes - elle atteint au maximum 30 % - pour les réductions temporelles les plus fortes.

Les résultats obtenus pour le contexte /a/ seront examinés en un premier temps.

Par la suite, nous analyserons les écarts entre les résultats correspondant aux différents contextes vocaliques en nous basant sur les données recueillies pour les tenues les plus brèves -rapport temporel = .1. Ce n'est qu'à ce niveau que certaines différences apparaissent à l'intérieur du groupe constitué par les environnements vocaliques mi-fermés et fermés

1. /apa/, /ata/ et /aka/.

L'examen par un plan d'analyse de la variance à 3 facteurs (consonne, durée de le tenue, sujets) montre que les effets "consonne", "durée de la tenue" et l'interaction "durée x consonne" sont significatifs à .001.

La poursuite de l'analyse par contrastes S de SCHEFFE (1959) ($p < .05$) indique que :

- en regroupant les résultats sur l'ensemble des durées, les consonnes se répartissent en 2 groupes /ata/ et /aka/ d'une part, /apa/ d'autre part.
- que la différence entre les résultats de ces deux groupes est significative pour les rapports temporels de .5, . 4 et . 3 mais non pour . 1.
- que la décroissance du taux de réponses sourdes devient significative à partir de . 5 pour /ata/ et /aka/, à partir de . 3 pour /apa/.

Pour chaque consonne, la réduction temporelle de la tenue n'introduit un changement significatif dans les réponses des sujets que lorsqu'elle est conduite au-delà des valeurs qui caractérisent les occlusives voisées (voir figure 2). Cette divergence entre mesures acoustiques et résultats de perception est probablement due à la présence d'autres indices de non-voisement dans les séquences VCV.

La persistance du caractère non-voisé de /p/ peut être attribuée à la transition de F_I de la voyelle précédente qui constitue un indice de non-voisement particulièrement marqué pour la séquence /apa/ (WAJSKOP et SWEERTS, 1973).

2. Influence de l'environnement vocalique.

La comparaison entre contextes vocaliques fermé et ouvert fait apparaître 2 différences* au niveau des corrélats acoustiques de l'opposition voisé/non-voisé :

1° le recouvrement entre les distributions des rapports temporels tenue/voyelle précédente, mesurés sur des occlusives voisées et sourdes, augmente progressivement de /a/ vers /i/ (voir figure 2).

2° le délai d'établissement du voisement (VOT) s'allonge lorsque le degré de fermeture de la voyelle augmente (voir figure 3).

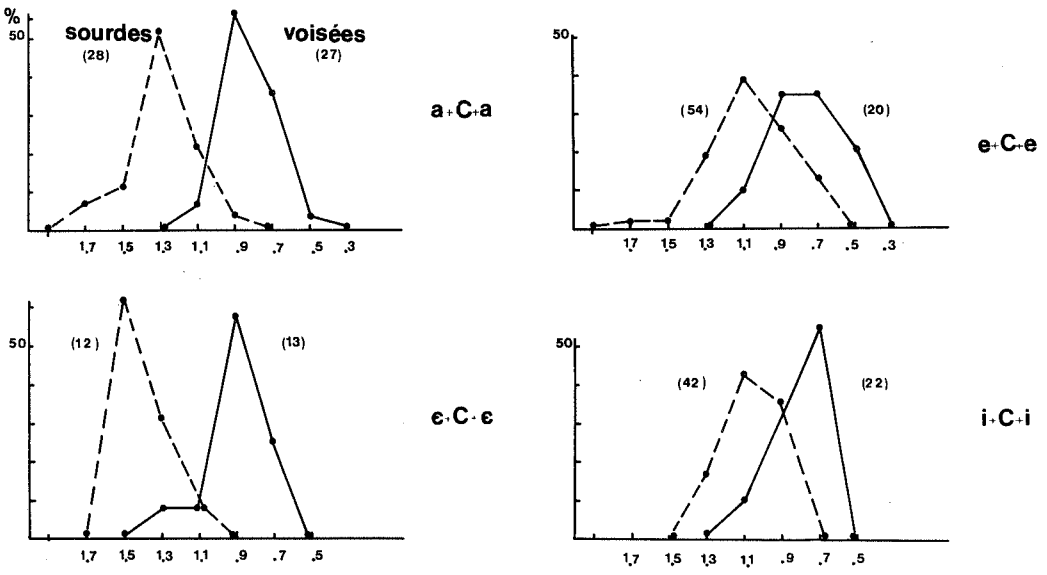


Figure 2 : Mesures acoustiques des rapports temporels tenue/voyelle précédente. Les rapports sont indiqués en abcisse. En ordonnée, nous avons mis les pourcentages de cas observés pour chaque rapport. Les nombres repris entre parenthèses correspondent aux effectifs de chaque échantillon. Les mesures ont été effectuées sur des séquences VCV prononcées par un même locuteur.

---

* Les mêmes différences avaient été observées par E. FISCHER-JØRGENSEN (1968) chez un locuteur francophone.

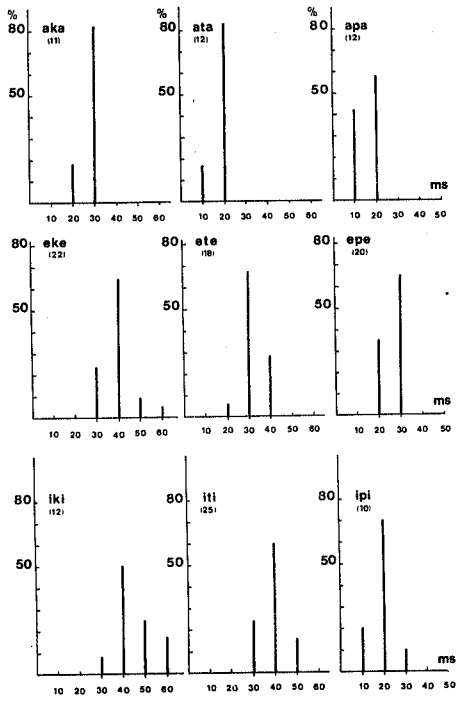


Figure 3 : Mesures acoustiques des délais d'établissement du voisement (VOT). En abcisse: VOT des occlusives sourdes en ms. En ordonnée : pourcentages de cas observés. Les nombres repris entre parenthèses correspondent aux effectifs de chaque échantillon. Les mesures ont été effectuées sur des séquences VCV prononcées par un même locuteur.

Au niveau acoustique, la dégradation du pouvoir distinctif de la durée de la tenue semble être compensée par l'intervention d'un autre indice - le délai d'établissement du voisement.

Qu'en est-il au niveau des résultats de perception ?

Afin de répondre à cette question les résultats obtenus pour les séquences VCV, au rapport temporel tenue/voyelle précédente = .10, ont été mis en relation avec leur VOT (voir figure 4).

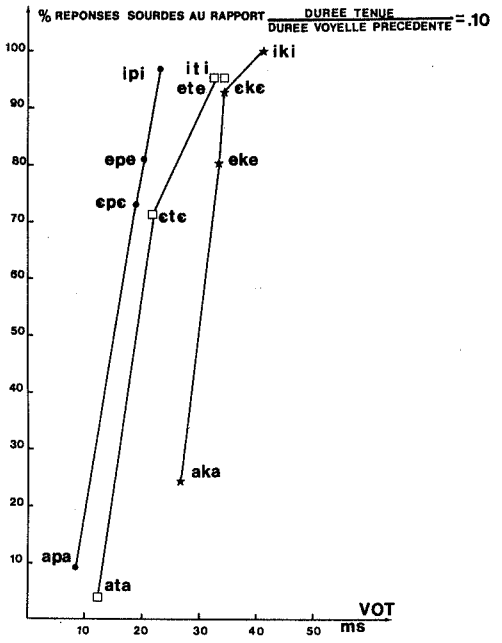


Figure 4 : Résultats de l'expérience d'identification - pour le rapport temporel tenue/voyelle précédente = .10 - en fonction du VOT des stimuli.

L'évolution du taux de réponses sourdes en fonction du VOT doit être considérée séparément pour chaque consonne. Les études menées sur des occlusives anglaises notamment (ABRAMSON et LISKER, 1970) montrent que la frontière perceptive dépend du point d'articulation : VOT plus bref pour la labiale, intermédiaire pour la dentale et plus long pour la vélaire.

Bien que nous ne disposions pas d'une relation directe entre la valeur acoustique du VOT et ses effets perceptifs, cet indice semble être à l'origine de la persistance du caractère non-voisé des occlusives comprises dans les environnements vocaliques fermés et mi-fermés.

En effet, pour chaque consonne, le pourcentage de réponses sourdes croît de manière plus ou moins régulière en fonction de l'allongement du VOT. En d'autres termes, l'influence de la durée de la tenue est d'autant plus faible que le VOT constitue une marque acoustique plus forte du caractère sourd de la consonne.

IV. CONCLUSION.

Malgré ses limites, la relation qui a été établie entre les mesures acoustiques du VOT et les scores d'identification montre qu'il s'agit d'un indice important pour la perception du non-voisement des occlusives comprises dans des environnements vocaliques fermés et mi-fermés.

Des travaux portant sur la manipulation directe du VOT apporteront probablement une confirmation à ce sujet.

Sans préjuger des résultats qui seront obtenus pour d'autres indices, et dans d'autres environnements vocaliques, il est d'ores et déjà certain que la durée de la tenue et le délai d'établissement du voisement constituent, chacun dans des contextes différents, des indices essentiels pour la perception du non-voisement des occlusives intervocaliques.

BIBLIOGRAPHIE.

Fant, G. - Sound, Features and Perception, S.T.L. - Q.P.S.R. , 2-3, Stockholm, 1967.

Fischer-Jørgensen, E. - Voicing, Tenseness and Aspiration in Stop Consonants, with Special Reference to French and Danish, ARIPUC, 3, 1968, 63-114.

Lisker, L. and Abramson, A.S. - The voicing dimension : some experiments in comparative phonetics, Proceedings of the Sixth International Congress of Phonetic Sciences , Prague, 1967, 563-567.

Ohman, S.E.G. - Coarticulation in VCV Utterances : Spectrographic Measurements, Journal of the Acoustical Society of America , 39, 1966, 151-168.

Scheffé, H. - The Analysis of Variance. New York, J. Wiley, 1959, 66-72.

Serniclaes, W. - La simultanéité des indices dans la perception du voisement des occlusives, Actes des 4èmes Journées d'Etude du Groupe de la Communication Parlée, GALF, Bruxelles, 1973, 359-369.

Wajskop, M. and Sweerts, J. - Voicing cues in oral stop consonants. Journal of Phonetics, 1, 1973, 121-130.

LES CARACTERISTIQUES INTRINSEQUES DE LA FREQUENCE LARYNGIENNE : PRODUCTION, REALISATION ET PERCEPTION

Louis-Jean BOE
Institut de Phonétique
Grenoble

et

Danièle LARREUR
C.N.E.T.
Lannion

Résumé

D'une façon générale les faits prosodiques sont réalisés entre autres par l'intermédiaire de trois variables : la fréquence laryngienne, l'intensité globale et la durée. L'évolution de la première en fonction du temps revêt en français une grande importance perceptuelle et peut être considérée comme la superposition de phénomènes extrinsèques et intrinsèques : les uns ont une valeur conventionnelle, les autres sont uniquement liés au processus articulatoire de production. En français, la hauteur et ses variations, liées à la réalisation des voyelles et des consonnes n'ont pas de valeur distinctive. Ces caractéristiques fréquentielles peuvent s'expliquer par le couplage source laryngienne- conduit vocal, qui peut se présenter sous deux aspects différents liés au mode de production :

- la disposition articulatoire est telle que l'on peut considérer que la pression supra-glottique reste constante,
- la constriction supra-glottique est importante et la pression intra-glottique devient fonction du type d'articulation considéré.

Cette différentiation correspond à l'opposition [+ consonnantique] [- consonnantique] .

Le but de cette étude est d'expliquer l'origine :

- des différences de hauteur que l'on relève entre les voyelles d'aperture différente,
- des variations caractéristiques de la fréquence laryngienne liées au consonnnes. On cherche également à spécifier ces dernières, dans le cas des logatomes ou de la parole continue, et à préciser leur importance perceptuelle.

Summary

Prosodic features are generally realised by means of three parameters : laryngeal frequency, global intensity and duration. The evolution of the first of them has a great perceptual importance in french and looks as a superposition of extrinsic and intrinsic phenomena : the former have a conventional value, the latter are only related to the articulatory process. In french the pitch and its variations do not have any distinctive value, as far as they solely depend on the vowel and consonant realisation. These frequency characteristics have an explanation in the coupling between the laryngeal source and the vocal tract. The coupling has two aspects, depending on the production manner :

- the articulatory disposition is in such a way, that the supra-glottic pressure remains constant,
- the supra-glottic constriction is important and the intra-glottic pressure becomes a function of the articulatory disposition itself.

This differentiation is related to the phonetic opposition [+ or - consonantal] .

The present study aims at explaining :

- the pitch differences in vowels of different apertures,
- the characteristic laryngeal frequency variations related to the consonants. An effort is made to specify these variations in the case of logatoms or continuous speech and to precise their importance on the level of perception.

LES CARACTERISTIQUES INTRINSEQUES DE LA FREQUENCE LARYNGIENNE : PRODUCTION, REALISATION ET PERCEPTION

Louis-Jean BOË
Institut de Phonétique
Grenoble

et

Danièle LARREUR
C.N.E.T.
Lannion

INTRODUCTION

L'évolution de la fréquence fondamentale du signal de la parole reflète la réalisation de trois phénomènes bien distincts : l'intonation de la phrase, l'accent de mot ou de syntagme, les caractéristiques intrinsèques. Ces dernières inhérentes au processus articulatoire de production, par définition ne participent pas directement au mécanisme du codage-décodage de la chaîne parlée. Si l'on prend comme hypothèse de départ la similarité de base des mécanismes articulatoires des locuteurs quelle que soit leur langue, on peut s'attendre à ce que ces caractéristiques soient universelles comme les données articulatoires dont elles dépendent. Si l'on opère des mesures sur la fréquence fondamentale, toutes choses égales par ailleurs (entourage, accent, position dans le schéma intonatif) on relève des différences qui ne sont pas l'effet du hasard entre les voyelles et les consonnes sonores et à l'intérieur de ces catégories, et il existe une influence des secondes sur les premières. Au cours de la réalisation d'une consonne sonore on peut noter une évolution de la fréquence fondamentale que l'on peut relier au mode de production.

Ces faits ont été étudiés systématiquement depuis les premières observations de CRANDALL[1] qui datent de 1925. Par rapport à ces travaux, cet exposé a pour but d'apporter un certain nombre de précisions sur le production, la réalisation et la perception des caractéristiques intrinsèques dans le domaine du français. Il reprend une série d'études[2-8] qui ont été effectuées à l'Institut de Phonétique de Grenoble, au Laboratoire de la Communication Parlée de l'ENSERG et au Département ETA du CNET à Lannion.

PRODUCTION

Nous plaçant dans le cadre de la théorie myo-élastique moderne, nous considérons que les caractéristiques intrinsèques de la fréquence laryngienne sont dûes à des effets de couplage acoustique entre la source vocale et sa charge, c'est-à-dire entre le larynx et les cavités supra-glottiques. Ce phénomène peut prendre deux aspects différents correspondant à la division en sons [± consonantiques][9].

. Pour expliquer les différences de hauteur entre les réalisations vocaliques diverses hypothèses ont été avancées :

- lors de la production des voyelles fermées les muscles de la langue sont plus tendus que pour les voyelles ouvertes. Cette différence de tension serait due aux écarts entre la position de repos et les dispositions correspondant à ces apertures. Cette tension des muscles de la langue affecterait aussi ceux du larynx[10,11].

- pour les voyelles fermées le larynx se déplacerait avec la langue vers le haut et ce mouvement entraînerait une augmentation de la tension des cordes vocales[12-14].

- les voyelles postérieures ouvertes sont produites avec un rétrécissement de la section du pharynx. L'écoulement de l'air au niveau de la glotte présente donc des différences qui seraient à l'origine des écarts de fréquence[15].

La première hypothèse n'a pas été confirmée par de récents travaux radiographiques[16], la deuxième est en contradiction avec certaines mesures E.M.G.[17,18]. quant à la dernière elle apporte vraisemblablement une solution partielle, mais son interprétation reste limitée à l'importance de la seule zone du pharynx.
Aussi avons-nous tenté d'établir une explication systématique; elle part de la constatation suivante : dans toutes les études sur la production des voyelles, la charge du larynx a été considérée comme négligeable par rapport à son impédance interne, bien que cela ne soit qu'une première approximation[19-22]. FLANAGAN par exemple, souligne:"pour la fréquence du premier formant et à son voisinage il peut y avoir une interaction entre la source et le conduit vocal et en fait elle se produit". A l'aide d'un analogue[23] , nous avons relevé, en fonction de la fréquence, l'impédance d'entrée du conduit vocal pour les dispositions articulatoires correspondant aux voyelles orales du français. Les variations de cette impédance peuvent être reliées aux fonctions de transfert correspondantes : pour les fréquences formantiques la partie réelle est maximale et la partie imaginaire nulle (en deçà elle est positive et au delà négative). Si l'on compare les valeurs de l'impédance d'entrée ainsi relevées aux valeurs de l'impédance interne du larynx qui ont été calculées[21,22,24-27] , on constate qu'elles peuvent ne pas être négligeables. Comme le spectre de la source décroit de 12 dB/octave, ce sont les voyelles à premier formant bas qui vont intervenir le plus sur la source. Pour mettre en évidence les effets de cette interaction nous avons couplé un circuit charge à un modèle analogique de source, le modèle de PAILLE[22-28] . Comme l'étude théorique de ce circuit[29-31] , le laissait prévoir, une impédance de charge correspondant aux voyelles à faible 1° formant (voyelles de faible aperture) a tendance à élever la fréquence d'oscillation de la source et il y a un effet cumulatif des parties réelle et imaginaire . C'est la voyelle [a] qui influence le moins la source. Elle peut donc servir de référence.

Pour que cette explication soit valable, il faut s'assurer que les autres paramètres qui interviennent dans le régime de vibration des cordes vocales gardent les mêmes valeurs pour les différentes voyelles. Cela semble bien être le cas pour la pression sub-glottique[32-34] et les différentes tensions musculaires au niveau du larynx[17,18,35] .

. Pour les consonnes les phénomènes ont été nettement établis : les couplages entre la source et une charge dont l'impédance est très élevée va provoquer une augmentation de la pression supra-glottique et donc une diminution de la pression intra-glottique. Cette variation a pour conséquence de réduire l'effet Bernouilli et donc d'abaisser le régime de vibration des cordes vocales[35]. Avec les occlusives la source va fonctionner dans des conditions limites, si bien qu'elle peut cesser d'osciller en fin de tenue[15]. La relation entre la fréquence laryngienne et la pression intra-glottique a été systématiquement étudiée[33-45] : elles sont liées par une fonction croissante, linéaire ou logarithmique selon les chercheurs. Quantitativement on peut noter des différences

qui peuvent s'expliquer en partie par la diversité des procédures utilisées. Pour une variation de 1cm d'H_2O on peut noter, en moyenne, une évolution de la fréquence laryngienne de 5 à 7 Hz ou d'un demi-ton. Grâce à un appareillage de plus en plus adapté[45], les variations de la pression supra-glottique ont été mesurées en amont de la constriction ou de l'occlusion[20,33,34,40,46-61]. Tous les résultats sont concordants et cela pour différentes langues, la pression supra-glottique passe par un maximum au cours de la réalisation. Pour les occlusives et les constrictives les pressions maximales atteintes sont assez voisines, alors qu'elles sont nettement plus faibles pour les liquides et les nasales. Pour les deux premières catégories, compte tenu de la relation relevée entre la pression intra-glottique et la fréquence laryngienne, il faut s'attendre à des variations de l'ordre de 7% à 20%, c'est en position intervocalique que celles-ci risquent d'être les plus marquées.

REALISATION

La fréquence laryngienne des production vocaliques a été mesurée pour différentes langues[10-12,15,62-69]; les résultats sont comparables : entre [a] d'une part, [i] et [u] d'autre part, les différences peuvent atteindre plus de 10%. De notre côté nous avons relevé la fréquence des trois voyelles cardinales (position accentuée) avec un corpus de 132 logatomes CVCVC (entourage sonore) réalisé par 6 locuteurs français (3 hommes et 3 femmes). Si l'on prend [a] comme voyelle de référence, on note pour [i] et [u] + 14% et + 16%.

Dans l'ensemble les caractéristiques intrinsèques des consonnes ont été moins étudiées[1,11,15,68-70]. Nous avons utilisé ce même corpus, les résultats peuvent être regroupés en fonction des traits de mode et de lieu d'articulation. Les évolutions correspondent à celles que l'on pouvait prévoir à partir des relevés des variations de la pression supra-glottique. Pour les occlusives on observe, deux décrochements correspondant à l'ouverture et à la fermeture (en moyenne 9%), et pendant la tenue une variation exponentielle (environ 9% /10 cs.). Pour les constrictives la fréquence passe par un minimum, l'amplitude de déviation est de l'ordre de 12% à 21%. Pour les nasales et la latérale, la variation est aussi approximativement exponentielle avec des pentes de 7,5% à 14% /10 cs. Pour de la parole continue on constate un effacement des variations pour les nasales et, pour les autres sons, une uniformisation vers un tracé constrictif avec une inflexion moyenne de 15%. On peut relier cette évolution à la plus grande rapidité d'élocution qui modifie vraisemblablement la façon avec laquelle s'établit la pression supra-glottique.

PERCEPTION

Dans la mesure où les caractéristiques intrinsèques des voyelles ne sont pas réalisées sous forme de variations mais de différence absolue de hauteur, nous nous sommes limités à l'étude de l'influence de celles des consonnes sur l'intelligibilité et la qualité de la parole de synthèse, obtenue en l'occurrence par un vocodeur (2240 eb/s). Ces caractéristiques sont perceptibles, les études sur le seuil différentiel le laissent prévoir[71-73]. Mais, bien qu'elles aient été introduites à la synthèse[22,42,74-82], il ne semble pas que des études systématiques aient été faites pour préciser l'amélioration de la parole ainsi générée. On ne peut relever que quelques recherches assez limitées[83-85]. Nous avons effectué quatre séries de tests à partir de

logatomes et de chiffres. Le fondamental des consonnes sonores testées était réalisé constant, linéairement croissant ou présentant l'inflexion relevée à l'analyse. D'une façon générale les résultats confirment ceux qui ont été établis dans des études précédentes effectuées sur l'intelligibilité de la parole naturelle [86] ou synthétique [87-89]. Ils permettent d'autre part de préciser que restituées avec l'inflexion, sont améliorées la perception du trait de voisement des occlusives (+ 13% à 33% par rapport au contour constant), et la qualité des constrictives (par rapport aux deux autres contours).

1. CRANDALL, J.B. (1925), The Sounds of Speech. - B.S.T.J. 4, 586-626.

2. BOË, L.J. (1972), Etude de l'interaction source laryngienne-conduit vocal dans la détermination des caractéristiques intrinsèques des voyelles orales du français (fréquence laryngienne). - Bulletin de l'Institut de Phonétique de Grenoble 1, 25-43.

3. BOË, L.J.,(1973), Les faits prosodiques et la fréquence laryngienne. Approche théorique et expérimentale. - Bulletin d'Audiophonologie 2, 3-24.

4. BOË, L.J. (1973), Etude acoustique du couplage larynx-conduit vocal (fréquence laryngienne des productions vocaliques). - Revue d'Acoustique 27, 235-244.

5. LARREUR, D. (1972/73), Détermination des caractéristiques intrinsèques des consonnes voisées du français. Application à la synthèse. - Analyse et synthèse de la parole, CNET, Lannion 1, 21-37.

6. BOË, L.J. (1973), Etude de l'interaction source laryngienne-conduit vocal dans la détermination des caractéristiques intrinsèques des consonnes du français (fréquence laryngienne). Mesure de la durée. - Bulletin de l'Institut de Phonétique de Grenoble 2 (à paraître).

7. LARREUR, D. et BOË, L.J. (1973), Les caractéristiques intrinsèques des consonnes voisées du français dans la parole continue (fréquence laryngienne). - Bulletin de l'Institut de Phonétique de Grenoble,2, (à paraître).

8. LARREUR, D. et BOË, L.J. (1973), Etude de l'influence des variations de la fréquence laryngienne sur l'intelligibilité et la qualité des consonnes sonores générées par vocodeur. - Bulletin de l'Institut de Phonétique de Grenoble 2, (à paraître).

9. CHOMSKY & HALLE, M. (1968), The Sound Pattern of English. - Harper & Row Pub. New-York, trad. franç. Principes de Phonologie Générative, Ed. du Seuil, Paris, 1973.

10. TAYLOR, H.C. (1933), The Fundamental Pitch of English Vowels. - Journal of Experimental Psychology 16, 565-582.

11. HOUSE, A.S. & FAIRBANKS, G. (1953), The Influence of Consonant Environment upon the Secondary Acoustical Characteristics of Vowels. - J.A.S.A. 25, 105-113.

12. LADEFOGED, P. (1964), A Phonetic Study of West African Languages : an Auditory Instrumental Survey (West African Language Monograph Series 1). - Cambridge University Press, Cambridge.

13. WANG, W.S.Y. (1968), The Basis of Speech. - Project on Linguistic Analysis Report 4, Berkeley.

14. WANG, W.S.Y. (1968), The many uses of F_o .- Project on Linguistic Analysis Report W1-W35, Berkeley.

15. MOHR, B. (1971), Intrinsic Variations in Speech Signal. - Phonetica 23, 65-93.

16. NIELAGE, P. Mac (1969), A note on the Relations between Tongue Elevation and Glottal Elevation in Vowels. - Monthly Internal Memorandum, Phonology Laboratory. January, 86-90, University of California, Berkeley.

17. FAARBORG-ANDERSEN, K. (1957), Electromyographic Investigations of Intrinsic Laryngeal Muscles in Humans. - Acta Physiologica Scandinavia 41 (suppl. 140) 1-149.

18. FAABORG-ANDERSEN, K. (1965), Electromyography of Laryngeal Muscles in Humans. Technics and Results. - S. Karger, Bâle.

19. BERG, Jw. Van Den (1955), Calculations on a Model of the Vocal Tract for the Vowel /i/ (Meat) and on the Larynx. - J.A.S.A. 27, 332-338.

20. FANT, G. (1960), Acoustic Theory of Speech Production. - Mouton, The Hague, Paris.

21. FLANAGAN, J.L. (1965), Speech Analysis, Synthesis and Perception. - Springer-Verlag, Berlin - Heidelberg - New-York.

22. PAILLE, J. (1971), Contributions aux études sur la synthèse paramétrique de la parole. Synthétiseur à formants. Analogue de la source vocale. - Thèse d'Etat, Grenoble.

23. PULVERIC, F., VALET, J.Y., GAUDE, M. (1966), Simulateur analogique de l'appareil vocal. - ENSERG, Grenoble.

24. WEGEL, R.L. (1930), Theory of Vibration of the Larynx. - Bell System Technical Journal 9, 207-227.

25. BERG, Van Den Jw. - ZANTEMA, J.T. & DOORNENBAL, P. Jr. (1957), On the Air Resistance and the Bernouilli Effect of the Human Larynx. - J.A.S.A. 29, 626-631.

26. FLANAGAN, J.L. (1958), Some Properties of the Glottal Sound Sources. - J.S.H.R. 1, 99-116.

27. FLANAGAN, J.L. & LANDGRAF, L.L. (1968), Self Oscillating Source for Vocal-Tract Synthetizers. - IEEE Trans. Audio. and Elect. AU 16, 1, 57-64.

28. PAILLE, J. (1969), Source vocale pour synthétiseurs à formants. - Revue d'Acoustique 6, 111-114.

29. MOUSSIEGT, J. (1962), Oscillateurs à résistance négative et osscilateurs de relaxation. - Le Journal de la Physique et le Radium 23, 993-999.

30. MARCHANDEAU, R. (1962), Production d'oscillations par des diodes à effet tunnel. - Thèse - Grenoble.

31. MARCHANDEAU, R. & MOUSSIEGT, J. (1962), Variante simple de la méthode d'intégration graphique de l'équation aux oscillations de relaxation. - C.R. Acad. Sci. 254, 1236-1238.

32. STRENGER, F. (1958), Mesure de la pression sous-glottique, de la pression acoustique et de la durée de prononciation des différents sons du langage au cours de la phonation. - J.F. O.R.L. 7, 101-114.

33. LADEFOGED, P. (1963), Some Physiological Parameters in Speech. - Language and Speech 6, 109-119.

34. LADEFOGED, P. (1967), Three Areas of Experimental Phonetics. - Oxford University Press, London.

35. LIEBERMAN, P. (1967), Intonation, Perception, and Language. - Research Monograph 38, The MIT Press, Cambridge, Massachussetts.

36. BERG, Jw. Van Den (1956), Direct and Indirect Determination of the Mean Subglottic Pressure. - Fol. Phon. 8, 1-24.

37. BERG, Jw. Van Den (1957), Sub-glottic Pressures and Vibrations of the Vocal Folds. - Fol. Phon. 9, 65-71.

38. BERG, Jw. Van Den (1959), Données nouvelles sur la fonction laryngée. - J.F. O.R.L. 8, 103-111.

39. BERG, Jw. Van Den ((1960), Vocal Ligaments Versus Registers. - Current Problems in Phoniatrics and Logopedics 1, 19-34. - F. Trojan ed. S. Karger - Basel, New-York.

40. LADEFOGED, P. (1961), Physiological Studies of Speech. - STL - QPSR - 3, 16-21.

41. LADEFOGED, P. & KINNEY, N.P. Mac) (1963), Loudness, Sound Pressure and Sub-Glottal Pressure in Speech. - J.A.S.A. 35, 454-460.

42. ÖHMAN, S. & LINDQVIST, J. (1965), Analysis- by Synthesis of Prosodic Pitch Contours. - QPSR STL RIT, Stockholm 4, 1-6.

43. LIEBERMAN, P. & KNUDSON, R. & MEAD, J. (1969), Determination of the Rate of Change of Fundamental Frequency with Respect to Subglottal Air Pressure during Sustained Phonation. - J.A.S.A. 45, 1537-1543.

44. GLONE, R.E. Mc & SHIPP, T. (1970), Changes in Subglottal Air Pressure Associated with Changes of Fundamental Frequency. - J.A.S.A. 48, 118 (A).

45. HIXON, T.J., KLATT, D.H. & MEAD, J. (1971), Influence of Forced Transglottal Pressure Changes on Vocal Fundamental Frequency. - J.A.S.A. 49, 105.

46. FISHER-JØRGENSEN, E. & TYBJAERG-HANSEN, A. (1959), An Alectrical Manometer and its Use in Phonetic Research. - Phonetica 4, 43-53.

47. FISHER-JØRGENSEN, E. (1963), Beobachtungen über Zusammenhang Zwischen Stimmhaftigkeit und intraoralem Luftdruck. - Zeitschrift für Phonetik 16, 19-36.

48. ARKEBAUER, H.J. (1964), A Study of Intraoral Air Pressures Associated with Production of Selected Consonants. - Ph. D. Thesis, Univ. of Iowa.

49. ISSHIKI, N. & RINGEL, R. (1964), Air Flow during the Production of Selected Consonants. - J. S.H.R. 7, 233-244.

50. LISKER, L. (1965), Supraglottal Air Pressure in the Production of English Stops. - Status Rep. Speech Res. , Haskins Lab. 4, 3.1 - 3.15.

51. LISKER, L. (1966), Measuring Stop Closure Duration from Intraoral Pressure Records. - Status Rep. Speech. Res. Haskins Lab. 7/8, 5.1 - 5.6.

52. YAHAGIHARA, H. & HYDE, C. (1965/66), An Aerodynamic Study of Articulatory Mechanism in the Production of Bilabial Stop Consonants. - Studia Phonologica 4, 70.

53. MALECOT, A. (1966), The Effectiveness of the Intra-oral Air-Pressure-Pulse Parameters in Distinguishing between Stop Cognates. - Phonetica 14, 65-81.

54. SUBTELNY, J.D., WORTH, J.H. & SAKUDA, M. (1966), Intraoral Pressure and Rate of Flow during Speech. - J.S.H.R. 9, 498-518.

55. ARKEBAUER, H.J., HIXON, T.J. & HARDY, J.C. (1967), Peak Intraoral Air Pressures during Speech. - J.S.H.R. 10, 196-208.

56. MALECOT, A. (1968), The Force of Articulation of American Stops and Fricatives as a Function of Position. - Phonetica 18, 95-102.

57. NETSELL, R. (1969), Subglottal and Intraoral Air Pressures during the Intervocalic Contrast of /t/ and /d/. - Phonetica 20, 68-73.

58. AGNELLO, J.G. & GLONE, R.E. Mc. (1970), Differentiation of /p/ and /b/ from Spectrographic- Intraoral Air Pressure Comparison. - J.A.S.A. 48, 121 (A.).

59. HIKI, S., KOIKE, Y. & TAKAHASHI, H. (1970), Simultaneous Measurements of Subglottal and Supraglottal Pressure Variations. - J.A.S.A. 48, 118-119.

60. LISKER, L. (1970), Supraglottal Air Pressure in the Production of English Stops. - Language and Speech 13, 215-230.

61. LUBKER, J.F. - PARRIS, P.J. (1970), Simultaneous Measurements of Intraoral Pressure, Force of Labial Contact, and Labial Electromyographic Activity during Production of the Stop Consonant Cognates /p/ and /b/. - J.A.S.A. 47, 625-633.

62. BLACK, J.W. (1949), Natural Frequency, Duration, and Intensity of Vowels in Reading. - J.S.H.D. 14, 216-221.

63. PETERSON, G.E. & BARNEY, H.L. (1952), Control Methods used in a Study of the Vowels. - J.A.S.A. 24, 175-184.

64. LEHISTE, I. & PETERSON, G.E. (1961), Some Basic Considerations in the Analysis of Intonation. - J.A.S.A. 33, 419-425.

65. RAKOTOFIRINGA, H. (1968), Contributions à l'étude de la phonétique malgache. Hauteur, durée et intensité vocaliques efficaces. - Grenoble.

66. POTAPOVA, R.K. & BLOXINA, L.P. (1970), Prosodičeskie Xarakteristiki Reči (The Prosodic Characteristics of Speech). - Experimental Phonetics and Speech Psychology Laboratory, The Maurice Thorez Moscow State Pedagogical Institute for Foreign Languages, Moscow.

67. SAMARAS, M. (1972), Influence de l'entourage consonantique sur les variations de la fréquence laryngienne des voyelles du grec moderne. - Bulletin de l'Institut de Phonétique 1, 57-66.

68. MAACK, A. (1958), Regeln der Deutschen Silbenmelodie. - Phonetica 2, 199-219.

69. KIM, K. (1968), F_o Variations according to Consonantal Environments. - Monthly Internal Memorandum, September, 33-43, Phonology Laboratory. University of California, Berkeley.

70. MOHR, B. (1968), Intrinsic Fundamental Frequencies. IV : Voiced Consonants. - Monthly Internal Memorandum, September, 17-22, Phonology Laboratory. University of California, Berkeley.

71. FLANAGAN, J.L. & SASLOW, M.G. (1958), Pitch Discrimination for Synthetic Vowels. - J.A.S.A. 30, 435-442.

72. ROSENBERG, A.E. (1968), Effect of Pitch Averaging on the Quality of Natural Vowels. - J.A.S.A. 44, 1592-1595.

73. KLATT, D.H. (1973), Discrimination of Fundamental Frequency Contours in Synthetic Speech : Implication for Models of Pitch Perception. - J.A.S.A. 53, 8-16.

74. MATTINGLY, I.G. (1966), Synthesis by Rule of Prosodic Features. - Language and Speech, 9, 1-13.

75. ICHIKAWA, A. & NAKATA, K. (1968), Speech Synthesis by Rule. - Proc. 6th Int. Cong. Acoust. Paper B - 5 - 6.

76. ÖHMAN, S.E.G. (1968), A Model of Word and Sentence Intonation. - STL - QPSR 2-3, 6-11.

77. RABINER, L. (1968), Speech Synthesis by Rule. - B.S.T.J. 47, 17-37.

78. RABINER, L.R. & LEVITT, H. (1968), New Results in Speech Synthesis by Rule. - Proc. 6th. Int. Congr. Acoustic. Paper B - 5 - 14.

79. NEMETH, A. (1970), La Synthèse par règles de la parole. - 1° journées d'études sur la parole. Grenoble - Groupe de la Communication Parlée du GALF, 53-61.

80. PONCIN, J. (1970), Etude d'un système de synthèse de messages vocaux. - Annales des Télécommunications 11/12, 405-418.

81. NEMETH, A. (1971), Synthèse par règles de la parole à l'aide d'un vocodeur programmé avec sortie en modultation par impulsions codées. - 7th Int. Cong. Acoust. Budapest Paper 24 C2.

82. VAISSIERE, J. (1971), Contribution à la synthèse par règles du français. - Thèse de 3e Cycle - Université des Langues et Lettres de Grenoble.

83. SATO, T. (1958), On the Differences in Time Structures of Voiced and Unvoiced Stop Consonants. - J.A.S.J. 14, 117.

84. CHISTOVICH, L.A. (1969), Variation of the Fundamental Voice Pitch as a Discriminatory Cue for Consonants. - Soviet Physics - Acoustics 14, 372-378. Translated from : Akusticheskii Zhurnal 14, 449-456 (1968).

85. HAGGARD, M. & AMBLER, S. & CALLOW, M. (1970), Pitch as a Voicing Cue. - J.A.S.A. 47, 613-617.

86. MILLER, G.A. & NICELY, P.E. (1955), An Analysis of Perceptual Confusions among some English Consonants. - J.A.S.A. 27, 338-352.

87. VOIERS, W.D. (1968), The Present State of Digital Vocoder Technique : a Diagnostic Evaluation. - IEEE Trans. AU - 16, 275-279.

88. SMITH, P.S. (1969), Perception of Vocoder Speech Processed by Pattern Matching. - J.A.S.A. 46, 1562-1571.

89. PECKELS, J.P. & ROSSI, M. (1971), Le test de diagnostic par paires minimales. Adaptation au français du "Diagnostic Rhyme Test"de W.D. VOIERS. Compte rendu des 2° Journées d'études sur la Parole. Bd, Be, Bf, Bg. (Aix). Groupe de la Communication Parlée du GALF. - Revue d'Acoustique 1973, 27, 245-262.

D. ROSTOLLAND et C. PARANT
Laboratoire de Physiologie du Travail
du C.N.A.M. et du C.N.R.S. (Paris)

INFLUENCE DE L'INTENSITE SONORE DE LA VOIX SUR LA DUREE DES VOYELLES ET DES CONSONNES

On a étudié l'influence de l'"effort vocal" sur la durée des sons du langage. Un même matériel verbal (listes de mots dissyllabiques) a été prononcé suivant deux modes d'émission (voix parlée, voix criée) par trois locuteurs français. Pour chaque mot, de structure C.V.C.V., on mesure la durée des voyelles et celles de la deuxième consonne. L'augmentation de l'intensité sonore de la voix entraîne une légère diminution de la durée des consonnes et une forte augmentation de celle des voyelles. Ce résultat est discuté en relation avec les tests d'intelligibilité de la voix criée dans le bruit.

INFLUENCE OF VOICE LOUDNESS ON THE DURATION OF THE VOWELS AND CONSONANTS.

The effect of "vocal force" on the duration of speech sounds, has been studied. The same verbal material (word lists of two syllables) has been pronounced according to two modalities of utterance (speaking, shouting) by three french speakers. For each word, of the structure C V C V, the duration of the vowels and second consonant is measured. The augmentation of voice loundness involves two consequences : a small decrease of the consonants'duration, and a large increase of the vowels'duration. This result is discussed in relation to the intelligibility tests of the shouted voice in noise.

INFLUENCE DE L'INTENSITE SONORE DE LA VOIX SUR LA DUREE DES VOYELLES ET DES CONSONNES

par D. ROSTOLLAND et C. PARANT

Laboratoire de Physiologie du Travail du C.N.A.M. et du C.N.R.S. - Paris

Les caractéristiques physiques de la voix, si elles dépendent de l'âge et u sexe du locuteur, varient aussi de manière importante en fonction du niveau d'intensité adopté chez un locuteur donné. La fréquence fondamentale laryngée, par exemple, est très sensible à la variation de l'intensité émise, c'est-à-dire à l'effort vocal du locuteur. Avec un effort "maximum" (en voix criée), on observe une fréquence fondamentale qui dépasse le double de celle correspondant à un effort "normal" (voix parlée habituelle). Le fait d'élever la voix modifie, dans de larges limites, la distribution d'énergie acoustique suivant l'axe des fréquences et le long de l'axe du temps. On note, en particulier, les glissements de fréquence des trois premiers formants, ainsi que, d'autre part, les modifications d'amplitude et de durée relative des voyelles et des consonnes. Nous nous proposons, dans la présente étude, de préciser ce dernier point, c'est-à-dire de tester l'hypothèse d'une relation entre la fréquence fondamentale (ou l'énergie) et la durée des sons du langage.

Méthode et Résultats

Nous avons utilisé, comme matériel verbal, 24 mots dissyllabiques provenant des listes R. 5B (FOURNIER 1951). Ces mots ont été choisis de façon à avoir au moins 2 échantillons des consonnes fricatives, plosives, nasales, liquides et 5 échantillons des voyelles suivantes : i, e, ɛ, ɑ, o, ɔ̃ . Chaque mot a été prononcé, en voix parlée et en voix criée, par 3 locuteurs français. L'enregistrement est effectué en chambre sourde (t = 0,4 sec.) à 20 cm du microphone d'un sonomètre. Ce dernier, relié à un magnétophone Nagra fonctionnant à 19 cm/s, permet de contrôler l'intensité des deux types de voix : voix parlée à 80 dB et voix criée à 100 dB (valeur de crête). La visualisation des 144 mots, ainsi que la mesure des 432 intervalles de temps, est obtenue à l'aide d'un enregistreur U.V. (Honeywell, bande passante 0-10 KHz). Quelques sonagrammes ont dû être tirés, en bande large, afin d'évaluer les séparations entre voyelles et consonnes sonores. Pour chacun des mots, de structure CVCV et précédés de l'article "le", nous avons mesuré 3 intervalles de temps :

d_1 : durée du premier son vocalique ou de la première voyelle

d : durée entre la première et la deuxième voyelle ou durée de la deuxième consonne

d_2 : durée du deuxième son vocalique ou de la deuxième voyelle.

Nous n'avons pas pris en compte la durée de la première consonne, très difficile à déterminer surtout en voix criée. En effet, on ignore si l'intervalle compris entre l'article et la première voyelle doit être considéré comme la durée de la première consonne ou bien comme l'intervalle séparant deux mots.

Les résultats sont donnés par le tableau I. Pour un sujet et un type de voix donné la durée d_1 résulte d'une moyenne sur 24 mots (l'origine étant prise au début de la première voyelle). On trace alors la distribution dans le temps, puis l'on détermine d_1 à l'aide de la médiane. Les durées d et d_2 sont obtenues de la même manière, en effectuant deux changements d'origine successifs de valeur d_1 et d_1 + d.

Sujets	Voix parlée			Voix criée		
	d_1	d	d_2	d_1	d	d_2
A	130	80	250	160	70	250
B	130	110	210	180	80	390
C	140	120	230	170	90	430

Tableau I - Les nombres représentent des durées en ms.

On constate que l'intensité sonore de la voix a très peu d'effet sur la durée de la première voyelle et de la deuxième consonne. On note seulement pour d_1 une légère augmentation et pour d une légère diminution. Par contre, il semble que la durée d_2 soit sensible à l'augmentation du volume sonore de la voix puisque pour deux des sujets, elle est presque doublée en voix criée. Ce type de voix fait apparaître des différences entre sujets "parlant de la même façon". D'autre part, les histogrammes relatifs aux 18 durées montrent que la variabilité est très voisine pour les deux types de voix (valeurs extrêmes distantes de 90 ms en voix parlée et de 100 ms en voix criée).

Discussion et conclusion.

Tout se passe comme si le fait d'élever la voix conduisait à adopter une stratégie, indépendante de la signification des mots, qui consiste à augmenter la durée des voyelles et à diminuer légèrement celle des consonnes. En effet, la durée des voyelles passe de 180 ms à 265 ms et celle des consonnes de 105 ms à 80 ms, en moyenne. Cette stratégie correspond probablement à un compromis entre la nécessité d'accroître l'énergie de la "porteuse" et celle de ne pas détruire les consonnes, ou plutôt les transitions phonétiques, qui contiennent à elles seules presque toute l'information du message parlé. Une meilleure connaissance de l'influence de l'effort vocal sur les caractéristiques physiques de la voix est nécessaire pour interpréter les résultats, parfois paradoxaux, issus des tests d'intelligibilité de la parole en milieu bruyant (PICKETT, 1956, ROSTOLLAND et al. 1973). Cette étude peut également contribuer à une meilleure approche des problèmes de reconnaissance automatique de la parole, puisque le taux de réussite dépend, dans une certaine mesure, de la définition physique du signal de parole lui-même.

B I B L I O G R A P H I E

BORDONE C. and SACERDOTE G.G.,(1969) Some spectral properties of individual voices. Acustica 21-4

DUNN H.K. and WHITE S.D.,(1940) Statistical measurements on conversational speech. JASA 11-278

FAIRBANKS G. and MIRON M.S.,(1957) Effects of vocal effort upon the consonant-vowel ratio within the syllable. JASA 29-5

GARDE E., (1965) La voix - Presses Universitaires de France n° 627

HUSSON R., (1962) Le chant. Presses Universitaires de France n°997

IRWIN R.J. and MILLS A.W., (1965) Matching Loudness and vocal level : an experiment requiring no apparatus. British J. of P. 56-143

KAKUSHO O. and KATO K., (1968) Just discriminable change and matching range of acoustic parameters of vowels. Acustica 20-1

KLATT D.H., (1973) Interaction between two factors that influence vowel duration. JASA 54-4

KORN T.S., (1969) La dynamique de la parole dans les communications. Rev. Acoust. n° 3-4

LEIPP E., (1968) Structure physique et contenu sémantique de la parole. Rev. Acoust. n° 3-4

LIENARD J.S., (1972) Analyse, synthèse et reconnaissance automatique de la parole. Thèse. Faculté des Sciences. Paris

MALMBERG B., (1970) La phonétique. Presses Universitaires de France.

MIASNIKOV L.L. and MIASNIKOVA E.N., (1969) Ultrasonic speech components. Acustica 21-2

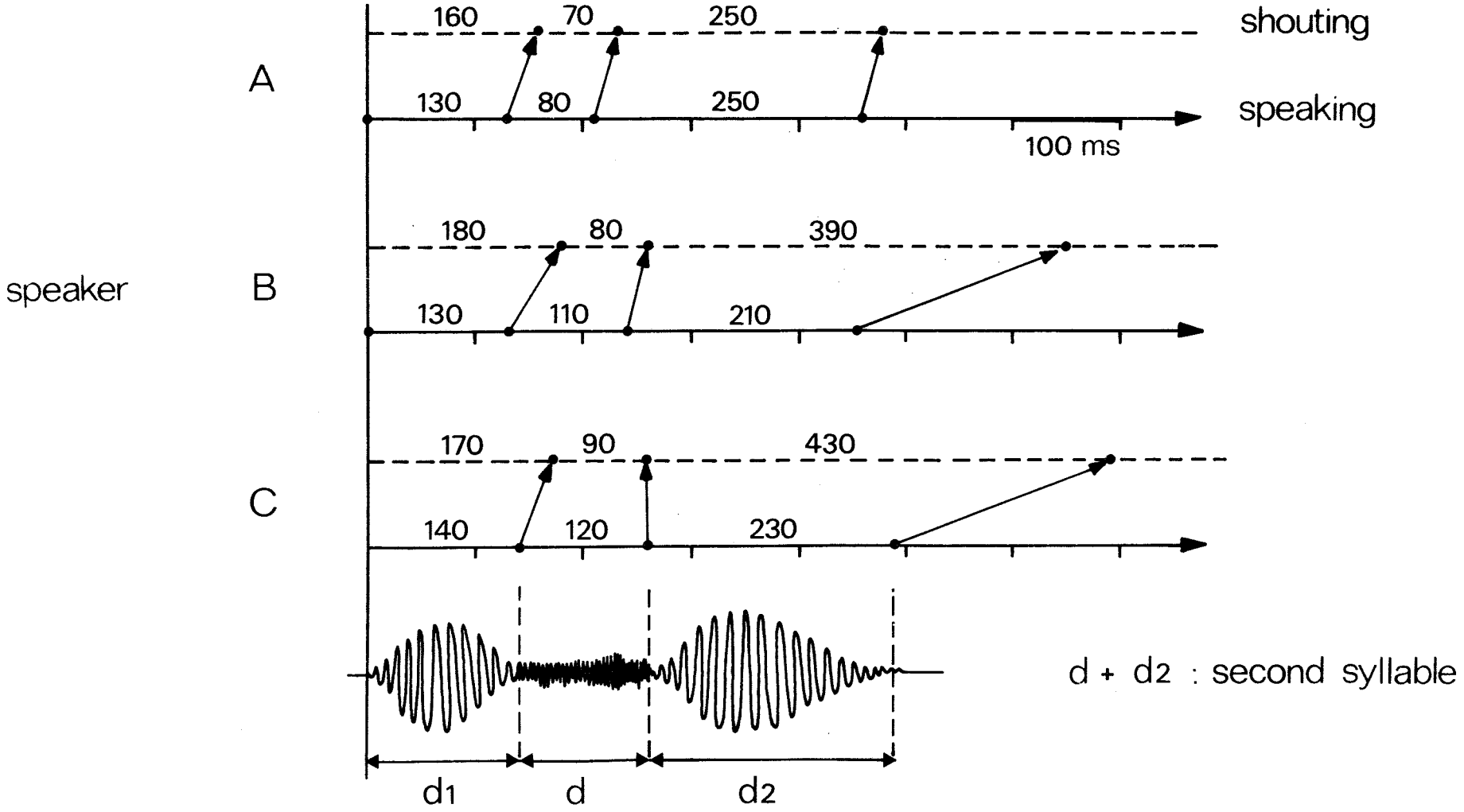
MOLES A., VALLANCIEN B., (1966) Phonétique et phonation. Masson. Paris

PICKETT J.M., (1956) Effects of vocal force on the intelligibility of speech sounds. JASA 28-5

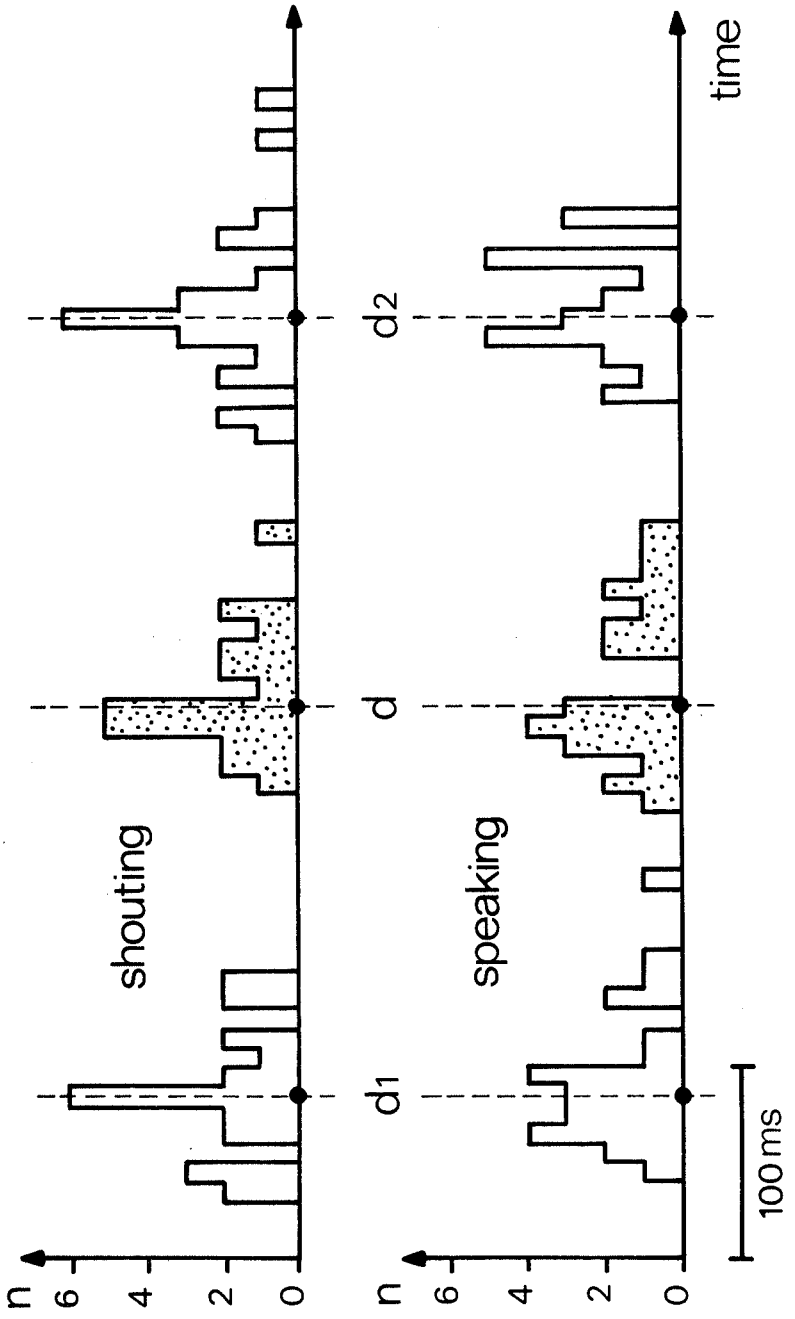
ROSTOLLAND D. and PARANT C., (1973) Distorsion and intelligibility of shouted voice. Symposium. Liège.

SJOGREN H., (1970) Objective measurements of speech level. Audiology 12.1

WAJSKOP M., (1970) Identification de voyelles en fonction de leur durée. Académie des Sciences. Prague.



DURATION ANALYSIS OF WORDS SEGMENTS



DISTRIBUTION OF WORD SEGMENT DURATION

speaker A

B. CAYLUX P. QUINTON
Université des Sciences Sociales de Grenoble
EQUIPE DE TRAITEMENT AUTOMATIQUE DES LANGUES

SYSTEME CONVERSATIONNEL D'ANALYSE SYNTAXIQUE DU FRANCAIS

Résumé

Nous présentons ici un système conversationnel d'analyse syntaxique de textes français. L'algorithme d'analyse est descendant et multiple, et indépendant de la grammaire qui est compilée de façon incrémentielle. Le système permet la recherche de structures complètes, ou la construction de structures partielles. Des résultats statistiques peuvent être extraits au niveau lexical ou syntaxique.

Summary

We present here an interactive system for syntactic analysis of French texts. The algorithm is bottom-up and grammar-independent. The grammar is incrementally compiled. The system makes it possible to find either complete or partial solutions and provides facilities for retrieving statistical information about the lexical and/or syntactical aspects of the text.

B. CAYLUX
P. QUINTON

Université des Sciences Sociales de Grenoble
EQUIPE DE TRAITEMENT AUTOMATIQUE DES LANGUES

SYSTEME CONVERSATIONNEL D'ANALYSE SYNTAXIQUE DU FRANCAIS

[1] L'objet de ce travail est de présenter un système d'analyse permettant de traiter des textes français.

A court terme, ce système sera utilisé par des linguistes, d'une part pour extraire automatiquement de différents corpus des renseignements (statistiques par exemple, concernant la fréquence des mots ou des syntagmes), d'autre part pour tester la limite de validité des hypothèses émises concernant la description structurelle de la langue.

A long terme, ces expérimentations devraient guider l'écriture d'une grammaire du français "performante", permettant de passer automatiquement d'un texte à un codage adapté à une application particulière, telle que la documentation automatique, les études stylistiques, la comparaison de textes ou l'analyse automatique du discours.

[2] L'utilisation du système se fait en quatre étapes :

- la description des catégories syntaxiques dont l'utilisateur entend se servir pour former les règles de reconnaissance.
- l'écriture des règles de reconnaissance qui devront s'appliquer sur le texte, ainsi que les conditions éventuelles portant sur les variables ou sur les codes syntaxiques attachés aux unités lexicales.
- l'utilisateur peut ensuite faire fonctionner ces règles sur un texte, ce qui lui permet d'en tester la validité.
- après mise au point de l'ensemble des règles, et selon l'application que l'utilisateur veut en faire, il précise la nature et la forme des résultats qu'il veut extraire du texte.

[3] Programme d'analyse

Le noyau du programme d'analyse est constitué par un analyseur syntaxique ascendant, multiple et général, acceptant les phrases d'une grammaire hors-contexte $G = (V_T, V_N, S, P)$, où

V_T est un ensemble de catégories lexicales,

V_N est un ensemble de catégories syntaxiques,

P est un ensemble de règles de la forme :

catégorie syntaxique → Ensemble régulier de chaînes de $(V_T \cup V_N)^*$

A chaque règle peut donc être associé un automate fini qui en reconnait la partie droite.

Le texte est lu de gauche à droite. Chaque mot est d'abord traduit par un programme de morphologie [1], en une ou plusieurs catégories lexicales munies de variables. L'analyseur syntaxique cherche à localiser dans la chaîne ainsi obtenue une partie droite de règle, c'est-à-dire une sous-chaîne reconnue par l'un des automates. Chaque sous-chaîne est alors remplacée par la catégorie syntaxique partie gauche, qui peut elle-même servir à reconnaitre une partie droite de niveau plus élevé. Le codage choisi permet de construire toutes les solutions en une seule lecture du texte. [2]

Afin d'éliminer certains calculs inutiles, on construit à la suite de l'analyse d'un mot l'ensemble des catégories successeurs possibles de ce mot.

Un exemple illustrant le déroulement de l'analyse est donné plus loin (voir [6].6)

[4] Compilation de la grammaire

La déclaration des règles de reconnaissance se fait en deux étapes :

1) la description des catégories syntaxiques : à chaque catégorie syntaxique ou lexicale est associé un ensemble de variables, elles-mêmes déterminées par l'ensemble de leurs valeurs possibles. (voir [6].1)
2) l'écriture des règles de reconnaissance qui devront s'appliquer sur le texte, ainsi que les conditions éventuelles portant sur les variables attachées aux unités lexicales. Une règles est constituée par la spécification d'une suite de catégories syntaxiques à localiser dans le texte (partie droite), et de la catégorie qui remplace cette suite après localisation (partie gauche).

La partie droite de la règle est une expression régulière construite à partir des opérateurs :

étoile (:*), concaténation (non marqué) et union (::), dans l'ordre décroissant de priorité. Les parenthèses permettent de modifier cet ordre. Chaque

symbole de l'expression régulière se présente sous la forme d'une étiquette suivie du nom de la catégorie à reconnaitre.

Les conditions qui doivent être satisfaites au moment de la reconnaissance sont décrites ensuite, et associées à un symbole par l'intermédiaire de l'étiquette ; elles peuvent porter sur l'accord de valeurs de plusieurs variables, ou sur une valeur particulière d'une variable. Le résultat (intersection des ensembles de valeurs de variable, dans le cas de l'accord) peut être affecté à la catégorie reconnue.

La compilation des règles produit des automates finis dont les arcs portent les instructions représentant les conditions et les actions.
La compilation de ces déclarations se déroule de façon conversationnelle : lors de la détection d'une erreur, le compilateur imprime un message, et l'utilisateur a alors la possibilité de faire immédiatement les corrections nécessaires. Par exemple, lors de la compilation d'une déclaration de catégorie, le compilateur vérifie que les variables associées à cette catégorie ont été déclarées au préalable. De même, lors de la déclaration d'une règle, il y a vérification de l'existence des catégories utilisées dans cette règle. Cette méthode de compilation incrémentielle permet la correction immédiate des erreurs, et par conséquent accélère la mise au point des déclarations.

[5] Moniteur d'exécution

Ce programme est prévu pour permettre à l'utilisateur de contrôler le déroulement du traitement, qui sera conversationnel.
Il permettra de définir, avant exécution, les options concernant le mode de déroulement, et la sortie des résultats.

Par la suite, en fonction de premiers résultats, nous comptons développer cette possibilité d'interaction afin d'augmenter la souplesse du système.

[6] Exemple

6.1 Commentaire déclaration des variables et de leurs valeurs ;

```
var gnr (masculin, féminin, neutre);
var nbr (singulier, pluriel);
var pers (un, deux, trois), temps (présent, futur, imparfait);
var mode (ind, cond, subj);
```

6.2 Commentaire déclaration des catégories lexicales article, substantif, adjectif, pronom objet, verbe et ponctuation ;

```
cat art (gnr, nbr);
cat subs (gnr, nbr), adj (gnr, nbr);
cat pro (gnr, nbr);
cat vrb (pers, nbr, temps, mode);
cat pct;
```

6.3 Commentaire déclaration des catégories syntaxiques groupe nominal, complément, sujet et phrase ;

```
cat gnom (gnr, nbr), comp (gnr, nbr);
cat suj (gnr, nbr);
cat phrase (temps, mode);
FIN.
```

6.4 Commentaire déclaration de la règle de reconnaissance de phrase ;

```
1 < p0 : phrase> ->  <p1 : suj> (<p2 : vrb> <p3 : comp> :: <p4 : pro> <p5 : vrb>)
                     <p6 : pct> ;
        p2 : acc (p2, p1) (nbr) ;
             prs (p2) = trois ;
        p5 : acc (p5, p1) (nbr) ;
             prs (p5) = trois ;
        p6 : temps et mode (p2 ou p5) => p0 ;
```

Commentaire déclaration des règles de reconnaissance du sujet et du complément ;

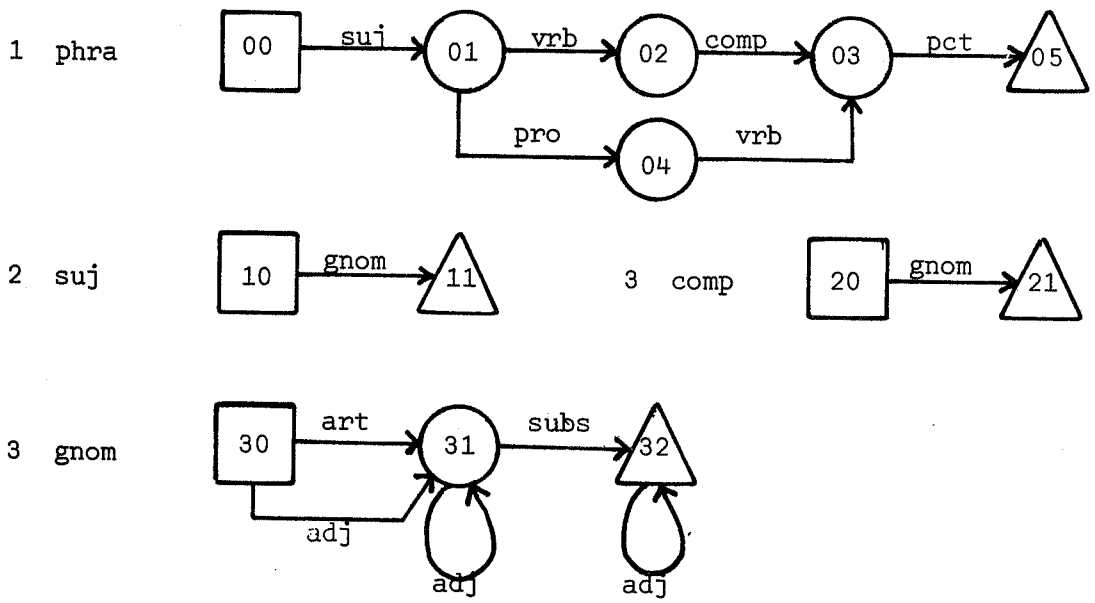
```
2 < s0 : suj> -> <s1 : gnom> ;
        s1 : gnr et nbr (s1) => s0 ;

3 < c0 : comp> -> <c1 : gnom> ;
        c1 : gnr et nbr (s1) => s0 ;
```

Commentaire déclaration de la règle de reconnaissance du groupe nominal ;

```
4 < g0 : gnom> -> (<g1 : art> :: <g2 : adj>) ¤<g3 : adj> <g4 : subs> ¤ <g5 : adj> ;
        g3 : acc (g3, g1) (gnr, nbr) => g0 ;
             acc (g3, g2) (gnr, nbr) => g0 ;
        g4 : acc (g4, g0) (gnr, nbr) => g0 ;
        g5 : acc (g5, g0) (gnr, nbr) => g0 ;
```

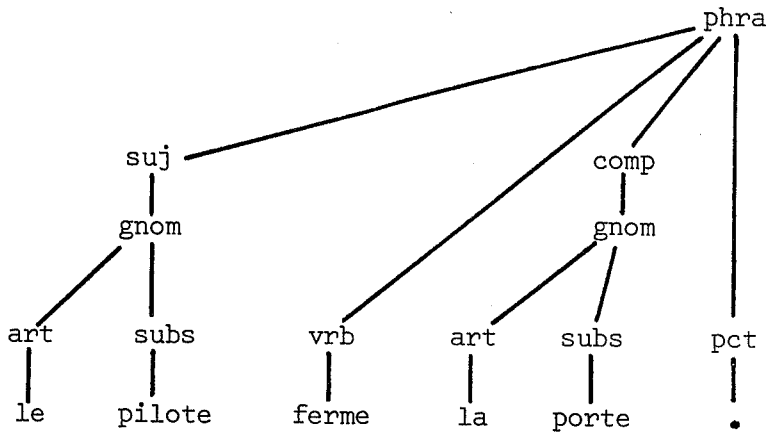
FIN

6.5. Automates (compte non tenu des actions)

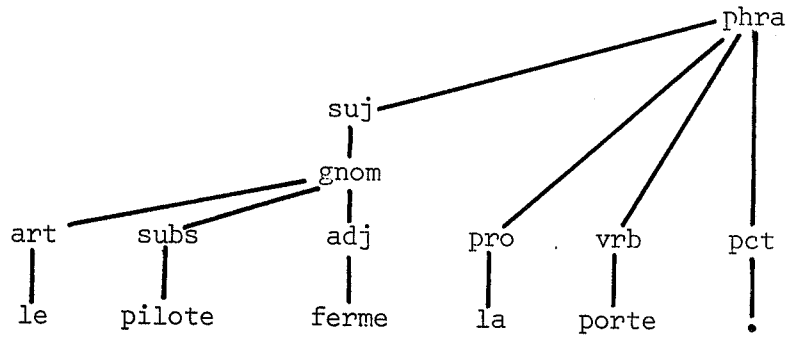


6.6 Exemple d'analyse de la phrase : "le pilote ferme la porte".

Structures



Solution 1



Solution 2

Trace de l'analyse

Les états sont accompagnés d'un indice donnant le numéro de leur constituant le plus à gauche. Un état provient soit de l'initialisation d'un automate (état 31 obtenu par *le*), soit de la transition d'un état déjà obtenu à l'aide d'une catégorie lexicale ou syntaxique (état 32, obtenu par 31 + *pilote* (subs)). Lorsqu'un état est final (ce qui est signalé par une *), on génère la catégorie syntaxique correspondante.

	Forme	*Résultat Morphologie*	*Etats et catégories construits*	*Successeurs possibles : phra, suj, gnom, art adj, subs*	*Commentaires Catégories initiales*
1	Le	pro	*		Début de la reconnaissance d'un groupe nominal.
		art	31, 1	adj, subs	
2	pilote	vrb	*		Suite du groupe nominal. L'état est final : on ajoute gnom, qui conduit à suj puis à la reconnaissance du début de phrase.
		subs .....>	32, 1 *		
			gnom, 1		
			11, 1 *		
			suj, 1		
			01, 1	vrb, pro	
3	ferme	adj .....>	32, 1 *		A cause de l'ambiguité de "ferme", on obtient un autre groupe nominal
			gnom, 1		
			11, 1 *		
			suj, 1		
			01, 1	vrb, pro	
		vrb .....>	02, 1	comp, gnom, art, adj, subs	ou la poursuite de la phrase
4	la	pro .....>	04, 1		Poursuite de la phrase ou début d'un complément
		art .....>	31, 4	adj, subs	
5	porte	vrb .....>	03, 1	pct	
		subs .....>	32, 4 *		
			gnom, 4		
			21, 4 *		
			comp, 4		
			03, 1	pct	
6		pct .....>	05, 1 *		
			phra, 1		

Bibliographie

(1) Editeur lexicographique pour les langues naturelles.
J. COURTIN, E. GRANDJEAN
Séminaire de Programmation de l'Université Scientifique et Médicale de Grenoble.

(2) An efficient Context-Free Parsing Algorithm.
J. EARLEY
CACM. vol. 13 / numéro 2 / Février 1970.

(3) Transitions Network Grammars for Natural Language Analysis.
W. WOODS
CACM / Octobre 1970.

(4) Regular expressions and State graphs for Automata.
Mc NAUGHTON, YAMADA
IRE. Transactions on electronic computers.

DONNEES STATISTIQUES SUR LA COMPOSITION PHONETIQUE DU FRANCAIS PARLE

M. MEPHAM
Université Laval - QUEBEC

Résumé provisoire

Une équipe de recherche sous la direction de M. Claude Rochette de l'Université Laval, à Québec, s'intéresse aux phénomènes de rencontre entre les phonèmes dans la chaîne parlée. Deux aspects importants de ces phénomènes sont les fréquences des différents types de rencontres, tels qu'ils se manifestent dans le discours, d'une part, et tels qu'ils se retrouvent dans le vocabulaire de la langue en cause, d'autre part.

Dans une première étape, quelque 40,000 entrées d'un dictionnaire de prononciation ont été dépouillées à l'aide de l'ordinateur. Le système de traitement, élaboré au Centre de traitement de l'information de l'Université Laval, a été conçu de façon à tirer le maximum d'information du corpus, qui comprenait, pour chaque mot la transcription phonétique, la syllabation orthographique et phonétique, et les catégories grammaticales. Nous disposons donc d'une masse de données considérables sur la composition phonétique et graphique du vocabulaire français, en forme de 20 tableaux et 10 listes; le tout comprenant plus d'un million de lignes d'impression.

Le contenu de ces tableaux et de ces listes est d'un grand intérêt à l'utilisation de contraintes phonétiques et phonologiques dans la reconnaissance de la parole.

Résumé et texte définitifs : voir le volume II des Actes

UN MODELE GENERATEUR DE MOTS PSEUDO-FRANCAIS RESPECTANT CERTAINES CONTRAINTES LINGUISTIQUES

J.S. LIENARD et C. CHOPPY
L.I.M.S.I. du C.N.R.S.

Résumé

Les résultats statistiques relatifs à la distribution des diphonèmes dans les mots du lexique sont utilisés pour créer des mots artificiels selon un processus aléatoire. Parmi ceux-ci la proportion de "mots accidentellement français" est d'autant plus élevée que les contraintes phonétiques, phonologiques ou morphologiques prises en compte sont voisines de celles existant dans la langue. L'intérêt d'un tel modèle est double : il permet d'une part, de créer des mots pseudo-français, dépourvus de signification, d'autre part d'étudier l'influence de "contraintes linguistiques virtuelles" sur la structure des mots du lexique.

A PSEUDO-FRENCH WORDS GENERATING MODEL SUPPORTING CERTAIN LINGUISTIC CONSTRAINTS

Summary

Statistical results related to the diphone distribution in lexical words are used to generate artificial words according to a random process. Among these, the "accidentally french words" percentage is all the higher as the given phonetic, phonological and morphological constraints are close to those existing in the langage. A such model is doubly interesting : on the one hand, it permits to create pseudo-french words without meaning and, on the other hand, to study the influence of "virtual linguistic constraints" on the lexical word's structure.

I - INTRODUCTION

Nous présentons ici un processus de création de mots possédant une "sonorité" française, mais ne figurant pas dans le dictionnaire de la langue française.

Cette étude a débuté en 1966 lorsque, pour des raisons tenant à la synthèse de la parole, nous avons effectué des comptages des digrammes phonétiques (diphonèmes) apparaissant dans des textes français écrits (1, 2). Il nous est alors apparu que la distribution des digrammes phonétiques obéissait à certaines normes statistiques, qui ne pouvaient se déduire des études classiques relatives à la distribution des phonèmes (3). Ainsi la pertinence de la notion de diphonème se trouvait-elle confirmée sur le plan de la linguistique. D'autres études, portant sur des bases plus larges (4) ou sur le français parlé (5) sont venues ensuite étayer notre point de vue.

Traitant, par ailleurs, de problèmes d'intelligibilité, nous avions besoin de mots sans signification, mais proches, phonétiquement, des mots courants du français (6). Rabelais, dans le cinquième livre de Pantagruel ("Comment furent les Dames Lanternes servies à soupper") donne quelques exemples savoureux :

- des badigonyeuses
- des happelourdes
- des cornicabotz
- des aucbares de mer etc ...

La tentation fut grande à ce moment, de créer de toutes pièces les mots dont nous avions besoin, en utilisant les tableaux de fréquence d'occurence des digrammes phonétiques (7). Le processus de création est décrit ci-dessous, avec diverses variantes, et des listes de mots pseudo-français sont données à la fin du texte.

Mais, plus peut-être que les mots pseudo-français eux-mêmes, la notion complémentaire de "mots accidentellement français" nous semble riche de conséquences. Elle permet, en effet, la définition d'un "taux de réussite" du processus, qui devient alors un véritable modèle phonétique actif du français, sur les paramètres duquel l'expérimentateur - car c'est bien d'une simulation expérimentale qu'il s'agit - peut jouer à l'infini.

Plusieurs personnes ont collaboré à cette recherche depuis son début, tant au niveau fondamental qu'au niveau de la programmation,

Melle M. CASTELLENGO, Mme A.M. LIENARD, Mrs E. LEIPP, D. TEIL, H. LUCOT et M. MLOUKA trouveront dans cet exposé un reflet de leur contribution, ainsi que nos remerciements.

II - STATISTIQUE ET CONTRAINTES PHONETIQUES

Par le terme de "contraintes phonétiques" on peut entendre l'influence des difficultés de réalisation de la langue sur son organisation phonétique. En particulier, le fonctionnement continu des organes phonatoires, leur inertie relative, les habitudes musculaires acquises dans l'enfance et sans doute bien d'autres facteurs font que tous les mouvements articulatoires ne présentent pas la même difficulté pour un individu donné. Si l'on admet que la majeure partie des individus d'un même ensemble linguistique utilisent les mêmes mouvements articulatoires pour réaliser les mêmes sons (ce qui n'est pas entièrement démontré, du moins dans le détail) ; si l'on admet aussi que la communication parlée cherche à transmettre le maximum d'information avec le minimum d'éléments phonétiques et dans un minimum de temps, alors on comprend l'existence d'une distribution stable des digrammes phonétiques traduisant un compromis entre la "difficulté d'articulation" de la langue, et son efficacité informative. Cette norme peut s'exprimer sous forme d'un tableau (fig. 1) ou sous forme d'une courbe de la fréquence d'occurence en fonction du rang (fig. 3) :

plus la courbe se rapproche de l'horizontale, et plus le système de digrammes choisi est "efficace" du point de vue de l'information, mais aussi plus il suppose un "entraînement" des locuteurs, visant à banaliser l'emploi des digrammes, c'est-à-dire à enchaîner un phonème à n'importe quel autre, indifféremment.

Il est intéressant d'examiner quelque peu les tableaux de fréquence d'occurence des digrammes phonétiques. On constate immédiatement que

a) selon que l'on distingue 30 à 36 phonèmes, en français, on devrait pouvoir construire 900 à 1296 digrammes phonétiques. Cependant la langue n'en utilise guère plus de 600.

b) la fréquence d'occurence d'un digramme est souvent sans rapport avec les fréquences d'occurence des symboles phonétiques

composants. Ainsi le digramme /də/ est-il environ 18 fois plus fréquent que l'on pourrait le croire à partir des fréquences de /d/ et /ə/ ; et /əd/ , par contre, est environ 8 fois moins fréquent que prévu.

Ces observations, choisies parmi bien d'autres, sont difficilement imputables au hasard, et mériteraient d'être mises en corrélation avec le fonctionnement des organes phonatoires et les habitudes de parole des adultes francophones.

Remarquons qu'il n'est pas indifférent de faire une étude statistique sur des textes choisis ou sur le discours parlé d'une part, sur un lexique d'autre part. Dans le premier cas certains mots sont répétés un grand nombre de fois (mots-clés d'un texte et mots-chevilles marquant la syntaxe), et "l'environnement phonétique" d'un mot (cest-à-dire les autres sons de la phrase) peut déterminer son choix ou son rejet par le locuteur. Rien de semblable dans le second cas : l'espace séparant un mot du suivant dans le lexique est une cloison étanche ; chaque mot n'apparaît qu'une fois et les désinences de la conjugaison sont omises, ainsi que les liaisons.

III - CREATION DE MOTS "ALEATOIRES" OU PSEUDO-FRANCAIS

Changeons maintenant de point de vue : au lieu de chercher les raisons de la distribution observée, considérons la comme une donnée. Nous pouvons fabriquer des mots artificiels respectant statistiquement cette distribution. Le processus est fort simple. Soit, par exemple, le tableau de la figure 1. Donnons nous un symbole phonétique de départ, soit /ɾ/ et faisons la somme cumulée de tous les nombres de la ligne, de la gauche vers la droite (fig. 2). La somme des fréquences d'occurence des digrammes commençant par /ɾ/ vaut 909, dernier nombre trouvé. En tirant au hasard un nombre compris entre 0 et 909, par exemple 150, on détermine le choix d'une colonne, c'est-à-dire d'un second symbole phonétique, soit /ɛ/ dans l'exemple de la figure 2. On peut recommencer l'opération à la ligne /ɛ/ et choisir ainsi un troisième symbole, etc... En répétant le processus n fois on fabrique un mot de n + 1 symboles phonétiques ; le premier symbole phonétique du mot peut être lui aussi tiré au hasard (cas du lexique) ou déterminé par une statistique des digrammes de liaison

des mots entre eux (cas du texte). En prenant certaines précautions statistiques (création d'un nombre de mots suffisamment grand pour que les comptages de digrammes aient un sens) on peut affirmer que les mots "aléatoires" présentent la même composition statistique que les mots utilisés pour faire les comptages initiaux (7).

Parmi les mots aléatoires, on trouve une certaine proportion de "mots accidentellement français". Toutes choses égales par ailleurs, cette proportion décroît avec la longueur phonétique des mots, selon une loi dont il faut chercher les raisons dans la combinatoire et dans la distribution des longueurs phonétiques des mots français. On peut avancer l'hypothèse suivante : plus les contraintes imposées lors du tirage aléatoire sont proches de celles qui régissent la langue, plus le taux de "mots accidentellement français" augmente, à longueur phonétique égale bien entendu, et en conservant un même lexique de référence. Donc ce taux représente quantitativement un "indice de réussite" du processus aléatoire.

Cette notion nous semble fondamentale. En effet, nous pouvons imposer, au moment du tirage, des contraintes tout-à-fait différentes de celles qui concernent les digrammes ; par exemple :

- ne pas prendre en compte les mots aléatoires comprenant plus de deux consonnes successives, ou ne comprenant que des voyelles, ou encore appliquer toute autre loi morphologique,
- sélectionner les mots aléatoires en fonction d'une statistique sur les trigrammes ou groupement d'ordre supérieur, sur les syllabes, sur certains radicaux ou désinences, sur les traits phonologiques, etc...

On peut, en somme, appliquer des contraintes virtuelles, de nature phonétique, phonologique ou morphologique, et savoir quantitativement si ces contraintes sont vérifiées dans la structure des mots français. L'organigramme de principe du processus est donné figure 4.

Notons encore le point suivant : parmi les mots aléatoires, pseudo-français ou accidentellement français, on peut avoir des répétitions. Sur 1000 mots aléatoires de 4 symboles phonétiques par exemple, on observera seulement 900 mots différant entre eux d'au moins un symbole phonétique. Plus les probabilités des différents digrammes possibles sont voisines les unes des autres, plus les tirages aléatoires peuvent être variés, et plus on trouve de mots différents,

Mais aussi plus on s'éloigne des structures phonétiques usuelles en français : on fabrique des mots plus "efficaces" du point de vue de la théorie de l'information, mais de plus en plus difficiles à prononcer, sinon impossibles. Nous retrouvons ici le principe de ZIPF.

IV - QUELQUES RESULTATS

Pour vérifier le bon fonctionnement du processus, nous avons effectué les expérimentations suivantes :

a) Statistique des digrammes phonétiques sur un lexique

Les 19181 mots d'un dictionnaire (Petit Larousse) ont été transcrits en phonétique, dans un code à 29 symboles, par le programme de traduction phonétique de D. TEIL .

Cette opération peut être critiquée, en ce que la transcription phonétique automatique, convenable pour les mots courants, fait quelquefois des erreurs. Celles-ci nous ont semblé suffisamment peu nombreuses pour ne pas compromettre l'expérimentation au niveau des principes.

Des comptages de digrammes ont été effectués ensuite, en différenciant les digrammes initial, médians et final de chaque mot. D'où trois tableaux statistiques différents. Un quatrième était obtenu sans différencier ces divers types de digrammes.

b) Création de mots aléatoires à partir de tableaux différenciés

Une liste de 1000 mots aléatoires a été éditée pour chaque longueur phonétique comprise entre 3 et 6 symboles phonétiques. Quelques mots pseudo-français sont donnés figure 5 ; pour la commodité, ces mots phonétiques sont accompagnés d'une transcription "orthographique" fantaisiste (nota : dans notre code phonétique, /E/ est une classe regroupant /œ/ , /ø/ et /ə/ ; /A/ regroupe /a/ et /ɑ/ ; /ɛ̃/ désigne aussi bien /œ̃/ ; /y/ désigne aussi /ɥ/ et /u/ désigne aussi /w/ ; enfin /ɲ/ est utilisé pour noter /ʃ/).

Le "taux de mots accidentellement français", relevé sur le dictionnaire initial, vaut 0,145, dans le cas des mots de 4 symboles phonétiques. Sur 1000 mots aléatoires de longueur 4 on en compte seulement 909 différents.

c) Création de mots aléatoires à partir du tableau unique (statistique indifférenciée)

Les digrammes de début et de fin de mot sont les plus affectés par cette contrainte virtuelle. Le taux de MAF4 passe à 0,068, et diminue donc de moitié. Mais l'uniformisation des probabilités fait que l'on trouve maintenant 970 mots différents sur 1000 mots aléatoires de longueur 4. Quelques mots pseudo-français sont donnés figure 6.

d) Création de mots aléatoires à partir de digrammes équiprobables et indifférenciés.

Dans ce cas, totalement artificiel, le taux de MAF4 tombe à 0,004 et l'on trouve 992 mots différents. Les mots pseudo-français prennent souvent une forme étrange, et deviennent difficilement prononçables (fig. 7). Ces mots seraient pourtant les plus efficaces, selon la théorie de l'information !

V - APPLICATIONS ET DEVELOPPEMENTS

Quelques applications ont été déjà évoquées plus haut. Dans le domaine des tests d'intelligibilité, il semble que les mots pseudo-français pourraient apporter une réponse satisfaisante au problème des mots phonétiquement équilibrés (8) et notamment remplacer les logatomes utilisés en télécommunications, qui sont en espéranto et ne représentent guère la langue française.

Nous avons assez dit ce que la notion de modèle phonétique actif de la langue nous semblait apporter de nouveau dans l'étude expérimentale des contraintes phonétiques, phonologiques et morphologiques. Nous pensons que cette démarche synthétique pourrait être étendue, approfondie, et pourrait compléter heureusement la démarche analytique des études statistiques habituelles.

On peut également envisager d'utiliser ce processus dans les études qui visent à créer des mots nouveaux : sous la pression du besoin, parce qu'une langue est vivante et doit faire face à de nouvelles applications, ou dans l'optique des "phonocodes" de J. DREYFUS-GRAF (9). Dans ce dernier cas on pourrait ne retenir que les mots "phonocodés" dont la structure phonétique soit d'une certaine manière voisine de celle d'une langue donnée.

VI - REFERENCES

1 - J.S. LIENARD - Le dictionnaire des éléments phonétiques et ses applications à la linguistique - Bulletin n° 22 bis du Groupe d'Acoustique Musicale de l'Université Paris VI, Juin 1966

2 - E. LEIPP, M. CASTELLENGO, J. SAPALY, J.S. LIENARD - Structure physique et contenu sémantique de la Parole - Colloque sur la Parole organisé par le GALF à Grenoble, Avril 1967
Revue d'Acoustique n° 3-4, décembre 1968

3 - P. GUIRAUD - Problèmes et méthodes de la statistique linguistique - Presses Universitaires de France, Paris 1960

4 - J.P. TUBACH - Etude des contraintes statistiques des groupements phonématiques - Rapport interne du Centre d'Etude pour la Traduction Automatique, Grenoble, 1970

5 - J.P. HATON, M. LAMOTTE - Etude statistique des phonèmes et diphonèmes dans le français parlé - Revue d'Acoustique n° 16, Décembre 1971

6 - E. LEIPP - Le problème de l'intelligibilité de la parole - Bulletin n° 37 du Groupe d'Acoustique Musicale de l'Université Paris VI, Novembre 1968

7 - J.S. LIENARD - Analyse, synthèse et reconnaissance automatique de la parole - Thèse d'Etat, Université Paris VI, Avril 1972

8 - T. TARNOCZY - A method for constructing phonetically balanced words - Symposium International sur l'Intelligibilité de la Parole, Liège, novembre 1973

9 - J. DREYFUS-GRAF - Codes phonétiques (phonocodes) et règles linguistiques - Comptes-rendus des 4e Journées d'Etude du groupe "Communication Parlée" du GALF. Bruxelles, Mai 1973

	A	E	i	o	ɔ	y	u	ɛ̃	ɑ̃	œ̃	e	ɛ	m	n	ʁ	l	j	ʃ	s	ʒ	z	f	v	p	b	t	d	k	g	Total ligne
	A	E	I	+	O	U	W	/	*	@	)	(	M	N	R	L	Y	X	S	J	Z	F	V	P	B	T	D	K	G	↓
A	0	1	0	0	0	0	0	0	0	0	1	0	32	50	151	123	29	12	99	40	21	18	27	33	46	90	36	49	14	885
E	1	0	2	0	0	0	0	0	0	0	0	0	38	22	80	26	0	2	4	1	40	4	16	14	5	29	6	13	2	304
I	27	13	0	2	20	3	0	76	11	8	57	42	31	86	66	42	32	6	103	12	40	32	29	20	18	86	28	70	13	988
+	0	0	0	0	0	0	0	0	0	0	0	0	1	0	2	1	0	1	4	0	1	2	2	0	1	8	5	1	1	40
O	2	2	0	0	0	0	0	0	0	0	1	2	33	64	98	60	0	8	21	9	15	9	6	22	15	40	15	29	11	472
U	3	3	19	0	0	0	0	0	2	0	5	7	22	11	60	42	0	2	16	3	11	2	2	10	10	16	8	12	2	281
W	66	1	10	0	0	0	0	0	0	6	3	2	0	0	38	12	0	4	13	2	2	5	8	7	2	14	3	2	0	210
/	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	18	5	0	6	4	12	8	20	14	6	3	100
*	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	6	52	0	0	6	3	16	15	87	32	11	8	258
@	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	14	3	0	7	2	15	2	20	13	12	5	99
)	6	0	2	0	5	0	0	0	4	0	0	1	23	20	55	23	0	11	13	7	19	12	15	27	12	38	18	35	14	373
(	0	1	0	0	0	0	0	0	0	0	0	0	6	32	137	42	15	4	62	2	12	5	3	0	1	41	3	34	12	421
H	66	12	55	3	33	16	15	6	53	3	37	17	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	324
N	42	15	63	5	21	15	6	5	8	1	51	27	0	0	0	0	25	0	0	0	0	0	0	0	0	0	0	0	0	290
R	112	65	160	7	63	31	24	13	37	12	104	54	24	25	0	5	0	7	24	11	0	6	10	8	14	39	26	11	6	909
L	56	24	100	5	36	30	11	8	27	7	53	38	3	0	0	0	0	0	3	0	0	2	2	2	1	10	1	3	1	437
Y	10	10	6	1	5	0	0	5	3	0	25	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	76
X	17	12	13	4	2	2	2	1	8	0	15	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	87
S	27	31	172	7	18	32	18	4	24	0	52	30	13	0	0	2	0	0	0	0	0	3	1	26	1	83	0	32	0	596
J	6	10	25	1	3	8	7	2	8	0	24	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	110
Z	13	5	34	1	10	6	1	5	6	1	29	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	122
F	20	2	39	2	16	11	9	4	5	1	12	17	0	0	25	18	0	0	0	0	0	0	0	0	0	0	0	0	0	190
V	27	13	42	3	12	3	11	0	14	3	17	29	0	0	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	189
P	65	12	45	4	41	18	16	5	11	5	30	27	0	0	60	32	0	0	5	0	0	0	0	0	0	12	0	0	0	395
B	30	4	30	4	16	14	22	4	4	1	12	14	0	0	38	43	0	0	5	1	0	0	0	0	0	1	0	0	0	261
T	63	40	126	7	29	45	21	7	27	10	105	40	1	0	117	0	0	0	1	0	0	0	0	0	0	0	0	0	0	652
D	16	16	76	3	17	16	7	3	15	2	86	13	1	0	26	0	0	0	0	1	0	0	0	0	0	0	0	0	0	306
K	86	8	13	4	60	33	23	56	11	5	12	11	0	1	40	24	0	0	39	0	0	0	0	0	0	29	0	0	0	464
G	28	2	5	4	10	5	8	1	4	1	4	4	1	0	38	13	0	0	0	0	8	0	0	0	0	0	0	0	0	145
Tot. col. →	808	3111	1047	75	428	298	213	215	294	93	744	414	238	320	1051	519	102	71	506	112	178	126	139	221	160	663	220	329	99	9999

Fig 1 - Tableau de fréquence d'occurrence des diphonèmes dans les mots d'un dictionnaire français (19181 mots ; 97370 occurrences, ramenées à 10 000). Les digrammes initial, médians et final du mot ne sont pas différenciés

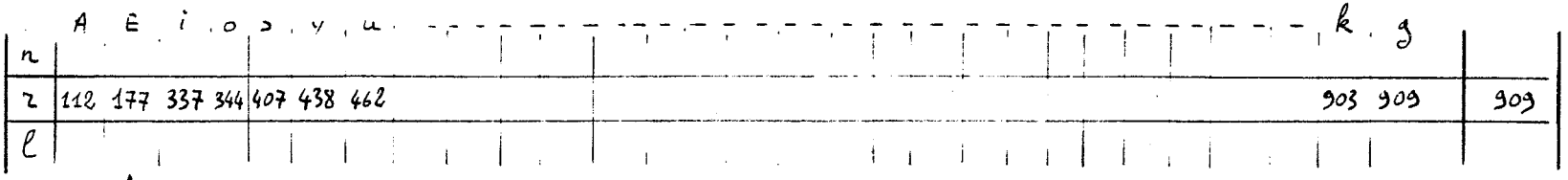


Fig 2 - Choix du 2e symbole phonétique d'un digramme (ici /ʁE/) par tirage d'un nombre aléatoire (ici 150) -

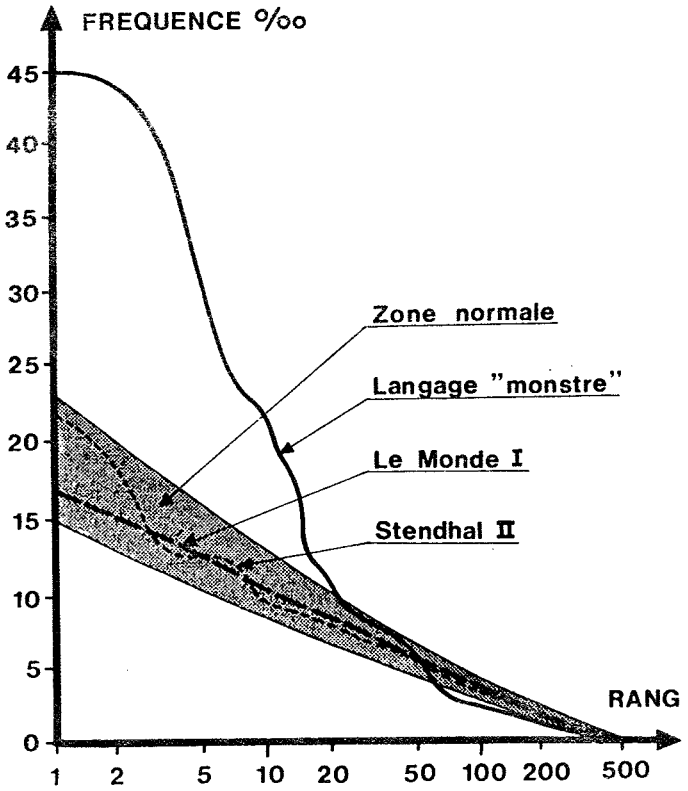


Fig 3 - Courbes de la fréquence d'occurrence en fonction du rang du diphonème. La quasi-totalité des textes étudiés se situent dans une zone restreinte (zone "normale"). Certains textes s'écartent notablement de cette zone, en particulier les textes techniques, les textes trop courts et certaines poésies.

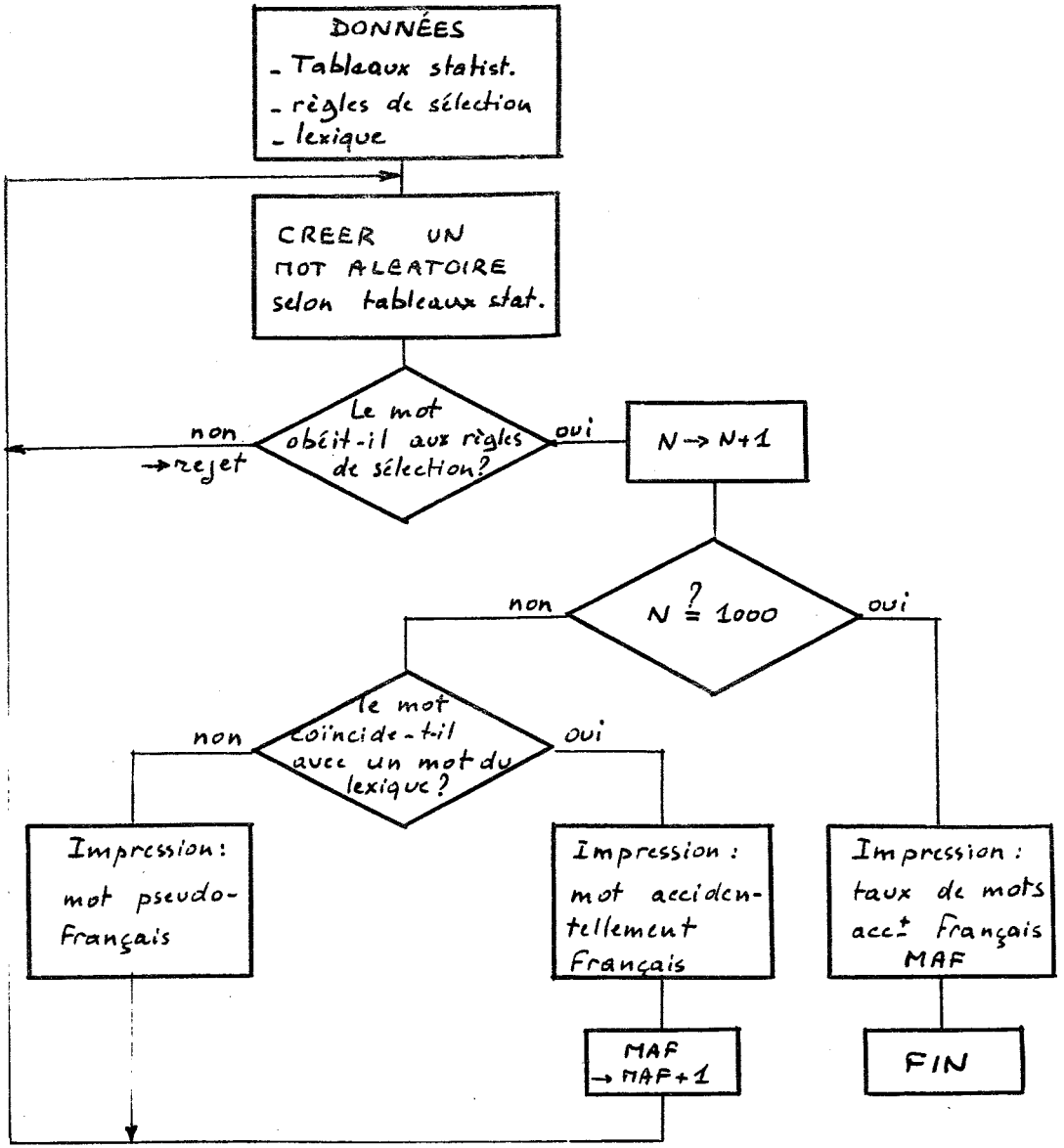


Fig 4 - Organigramme schématique du processus, pour une série de 1000 mots aléatoires. Chaque mot aléatoire est rangé dans l'une des deux catégories : pseudo-français, ou "accidentellement français".

/eʃe/	éché	/Abse/	absée	/ɛseze/	aissézé
/Ake/	aqué	/ɔsje/	ossié	/bɔkɔ̃t/	boconte
/tɛ̃k/	tinque	/ʒotA/	jauta	/AlAlɑ̃/	allalan
/reɑ̃/	réan	/ɔmɔg/	omogue	/ʒɔlɛl/	jolelle
/rEz/	reuse	/ʒipɔ̃/	gipon	/nɔlso/	nolseau
/ɛ̃ko/	inco	/Aksi/	haquecie	/bArin/	barine
/fle/	flé	/ɒlɔk/	ploque	/bɑ̃djɔ̃/	bandion
/dot/	daute	/myro/	mureau	/syima/	suima
/rer/	rér	/ɔsid/	osside	/nɔnɛr/	nonaire
/ɔbA/	oba	/sAje/	saillé	/dɛjɔr/	deillore

Fig 5 - Mots pseudo-français créés à partir de relevés statistiques sur les digrammes phonétiques du lexique, et retranscrits dans une "orthographe" fantaisiste. Les digrammes initial, médians et final sont différenciés.

/iekr/	ihécre
/ɛ̃djɛ/	indiais
/ɑ̃dir/	andire
/deit/	déite
/ArAd/	arade
/eril/	érile
/iɛlɛ/	iailet
/fAly/	fallu
/brEt/	brethe
/rmɔr/	rmor

Fig 6 - comme précédemment, mais à partir d'un tirage indifférencié.

/yipg/	uipgue
/fvug/	fvougue
/bɛdA/	bèda
/ɑ̃slo/	ancelot
/ɛfjv/	èffiu
/ɛ̃ndʃ/	intraduisible
/mubʃ/	moubche
/oron/	auraune
/ʒAsr/	jasre
/yjty/	intraduisible

Fig 7 - comme précédemment, mais à partir d'une distribution équiprobable des digrammes phonétiques.

Thème n° 2

SYNTHESE PAR REGLES

FREQUENCE FONDAMENTALE DES PHRASES DECLARATIVES EN FRANCAIS

J. VAISSIERE
Institut de Technologie du Massachusetts
Cambridge , USA

Résumé

L'étude en cours tente de définir approximativement le schéma intonatif utilisé par des locuteurs particuliers, à partir de la description phonémique et grammaticale des phrases. Ce résultat sera obtenu en ajustant des paramètres et en insérant des habitudes individuelles dans un ensemble de règles génératives. On a analysé le schéma intonatif de six locuteurs français prononçant dans leur langue maternelle des phrases isolées et un texte complet. Les résultats de l'analyse ont été comparés aux règles établies auparavant pour un locuteur professionnel. Les codes prosodiques utilisés par les sept locuteurs font l'objet d'une comparaison.

FUNDAMENTAL FREQUENCY IN DECLARATIVES IN FRENCH

Summary

Research in progress attempts to approximate the fundamental frequency pattern of individual speakers from the phonemic and grammatical description of the sentences. This will be done by adjusting parameters and inserting individual habits in the set of generative rules. The fundamental frequency pattern in isolated sentences and a text read by 6 native speakers of French were analysed. The results were compared with the rules previously found for one professional speaker. The prosodic codes used by the 7 speakers are compared.

FRÉQUENCE FONDAMENTALE DES PHRASES DECLARATIVES EN FRANCAIS

J.Vaissière
Institut de Technologie du Massachussetts
Cambridge,USA

Ce rapport est consacré à l'étude des corrélatifs acoustiques du caractère syntaxique de l'intonation dans des phrases déclaratives, prononcées par 7 locuteurs français. Dans une précédente étude*. nous avions établi, pour un locuteur professionnel, des règles sur les variations de la fréquence du fondamental selon la position, la longueur et la fonction grammaticale des mots dans la phrase. La présente étude a été effectuée à partir de l'analyse de phrases isolées et d'un texte de 142 mots (formé de phrases déclaratives), lus à des vitesses différentes par 6 locuteurs français (3 hommes et 3 femmes), non professionnels et non selectionnés. L'étude comparée des schémas intonatifs des 7 locuteurs (1+6) nous a permis de distinguer 2 types de règles: l'un correspond à des règles valables pour les 7 locuteurs (première partie), ou a des tendances communes (seconde partie), et l'autre correspond à des règles propres à un locuteur particulier ou à un groupe de locuteurs (troisième partie).

1. Règles communes:

A partir de nos observations, nous croyons pouvoir avancer l'hypothèse (non confirmée) que les caractéristiques communes des schémas intonatifs chez les 7 locuteurs correspondent à une structuration du flux de parole en un arbre syntaxique simplifié, à trois niveaux essentiels: la répétition de certaines variations caractéristiques de la fréquence du fondamental nous semblent en effet être essentiellement destinées à aider l'auditeur, premièrement à diviser le flux de parole en

unités de communication de base (ou Phrases), deuxièmement à lui signaler la division de ces phrases en Groupe de sens, et troisièmement à contribuer à l'interprétation de Mots successifs dans les groupes de sens. Le code commun aux 7 locuteurs est le suivant:

a) En fin de phrase déclarative, la fréquence du fondamental décroît rapidement sur les dernières syllabes et atteint au cours de la dernière la valeur la plus basse de la phrase. Si la phrase est courte (2 ou 3 syllabes), le schéma intonatif est constitué de cette seule chute. Dans une phrase plus longue (de 2 mots lexicaux au moins), la descente du fondamental est précédée pour les 7 locuteurs par une intonation montante sur la dernière syllabe de l'avant-dernier mot lexical (voir 2 exemples: "Je pars en vacances." et "Il étudie la dermatologie." sur la figure 1). Dans une phrase longue, la fin est marquée par l'un des deux schémas précédents, selon que les 2 derniers mots lexicaux sont séparés (Exemples: "...la dermatologie." et "...par la météo." sur la figure 1) ou non (Exemples: "...de la zone de travail." et "...au moulin de Melan." sur la figure 1) par une pause.

b) A un ou plusieurs mots liés par le sens (par exemple, le groupe sujet) ne correspond qu'un seul schéma de "groupe de sens" du point de vue prosodique: la figure 2 illustre les contours trouvés pour 4 de ces "groupes de sens" chez un locuteur particulier. Un groupe de sens prosodique comprend essentiellement 3 parties: une partie montante sur la ou les premières syllabes du groupe; deuxièmement, une zone où les impulsions positives ou négatives de la fréquence du fondamental relatives aux mots se combinent avec une pente de déclinaison générale (cette pente de déclinaison varie d'un locuteur à l'autre, et bien que son existence ne fasse aucun doute dans plusieurs langues, aucune explication claire de ce phénomène qui peut être d'ordre physiologique n'a été proposée);

et troisièmement, le groupe de sens est caracterisé par une remontée nette de la fréquence du fondamental au cours de la dernière syllabe (cf: la notion de continuation majeure en français de P. Delattre). Cette montée, suivie généralement d'une pause, marque l'existence d'une articulation importante, interne à la phrase, et la fin d'un groupe de sens.

c) L'existence de mots successifs sous-jacents aux groupes est corrélée 1- par au moins une impulsion positive de la fréquence du fondamental au niveau des mots lexicaux (par exemple, au niveau du nom commun, de l'adjectif, du verbe ...) et 2- par une impulsion négative ou par le maintien de la fréquence du fondamental à des valeurs basses au niveau des mots grammaticaux (par exemple, au niveau de l'article, de le préposition ou de l'auxiliaire ...). La figure 2 illustre quelques exmples de ces "impulsions" chez un des locuteurs: chaque mot lexical est caractérisé par une impulsion positive au niveau de sa première syllabe (l'hiver, retour, offensif, melon, Melun), alors que les mots grammaticaux sont prononcés avec des valeurs de fréquence du fondamental basses (un, de, des)(cf également les mots grammaticaux dans les phrases courtes ou les groupes de sens finals de phrases longues sur la figure 1).

2. Tendances communes:

Avant d'aborder le problème des divergences entre les locuteurs, nous allons essayer de résumer brièvement quelques-unes de leurs tendances communes:

a) Lorsque les locuteurs augmentent leur vitesse d'articulation, ils divisent la phrase par un moins grand nombre de pauses. La comparaison du nombre et de la position des pauses dans le même texte lu à des rythmes differents a montré que les premières pauses qui disparaissent se trouvent

aux jointures les moins importantes du point de vue grammatical (par exemple, entre le verbe et ses compléments)..A un rythme très rapide, seuls les signes de ponctuation ont été marqués par tous les locuteurs par une pause.

De même, la marque des jointures diminue en importance quantitative, et les mots successifs s'individualisent de façon moins claire. La figure 3b illustre un groupe de sens sujet prononcé à 2 rythmes différents par locuteur . On peut également observer, à un rythme rapide, un décalage vers la droite des pics de valeur du fondamental: comparez par exemple le déplacement de la valeur maximale de la première sur la seconde syllable du mot "intérêts" dans le groupe de sens sujet "dont les centres d'intérêts principaux...", prononcé à 2 rythmes différents, l'un jugé comme "normal" et l'autre comme "rapide" par le locuteur (cf: figure 4).

b) Il est aussi intéressant de noter que plus une unité d'un certain rang (phrase, groupe de sens ou mot) contient d'unités de rang immédiatement inférieur (groupes de sens, mots ou syllabes) plus les jointures entre les unités de rang inférieur sont clairement réalisées: les phrases très courtes ont même schéma intonatif; un groupe de sens formé de mots monosyllabiques sera plus difficile (voire impossible) à diviser en mots successifs à la simple vue des variations de la fréquence du fondamental qu'un groupe de sens comprenant des mots formés de plusieurs syllabes. Sur le plan de la perception, il est évident que des unités courtes pourraient être facilement décodées par le locuteur, même si elles étaient prononcés sans variations de fréquence fondamentale et sans rythme (même durée pour tous les phonèmes), alors que des unités plus longues exigent un effort de la part du locuteur (effort inconscient) pour diminuer celui de l'auditeur (effort dont l'auditeur devient

conscient quand il veut comprendre une voix mal codée du point de vue prosodique.

3. Variantes individuelles:

Les divergences entre les 7 locuteurs s'accentuent dans la démarcation des unités inférieures dans l'arbre syntaxique, c'est-à-dire dans la démarcation des mots. Par exemple, pour le locuteur dont quelques groupes de sens sont illustrés sur la figure 2, les mots internes aux groupes ont une intonation descendante et la jointure entre les mots lexicaux est essentiellement marquée par une impulsion positive au début du mot. La tendance dominante chez ce locuteur est un souci de régularité: chaque mot à l'intérieur du groupe a pas son intonation définie par les rapports grammaticaux du mot avec les autres mots du groupe.
Il n'en est pas de même chez tous les locuteurs (cf figure 3): une structuration secondaire s'établit à un niveau inférieur au groupe de sens et supérieur au mot. Lorsque 2 mots lexicaux se suivent dans un groupe, et sont très fortement liés par le sens (par exemple, un adjectif suivi du nom qu'il complète, ou un nom suivi de son adjectif, ou un expression connue -comme "centre d'intérêts"), le premier de ces 2 mots a une intonation descendante; dans le cas contraire, il a une intonation montante (cf: figures 2,3 et 4).
Néanmoins, chaque locuteur est constant avec lui-même et a ses courbes caractéristiques de fréquence du fondamental, dont les formes s'affirment lorsqu'il parle plus lentement: On pourrait classer les schémas en famille, comme il a été fait pour l'écriture: courbes arrondies (figure 3_c, par exemple), ou anguleuses (figure 3_d), généreuses (riches d'information syntaxique pour l'auditeur) ou ramassées...
Les variantes individuelles concernent également la position exacte des impulsions positives ou négatives dans les syllabes. Par exemple, l'im-

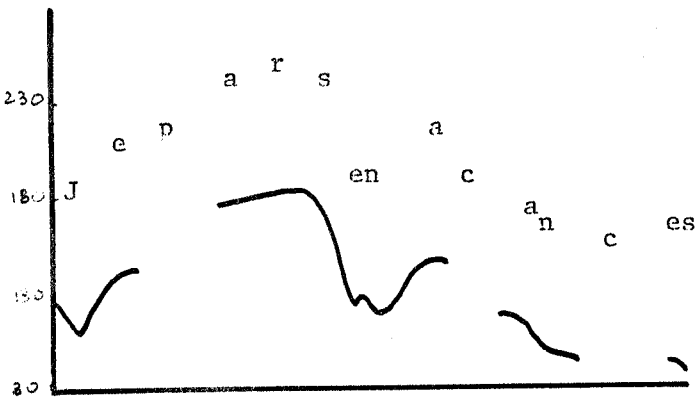
pulsion positive caractéristique de la fin d'un groupe de sens non final à la phrase peut débuter plus ou moins tôt dans la chaine des phonèmes selon les locuteurs et chacun a ses propres habitudes: par exemple, pour le locuteur de la figure 3c, elle débute au niveau de la dernière syllabe (voire même sur le dernier phonème de l'avant-dernière syllabe); pour le locuteur 3d, la fréquence du fondamental ne croît que sur la voyelle de la dernière syllabe.

Conclusion:

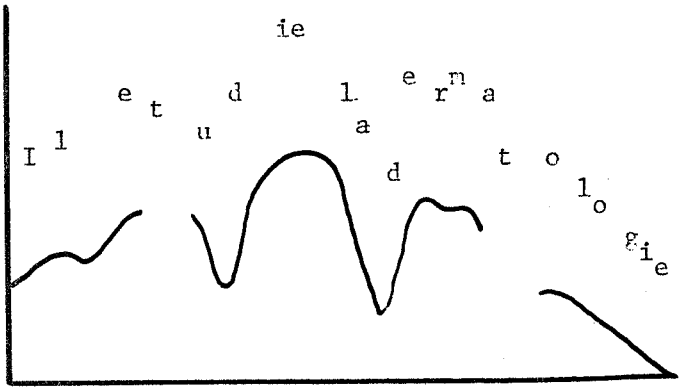
Il est á noter qu'en français, tous les facteurs prosodiques (intensité, pauses, rythme) semblent participer plus ou moins au rôle de structuration syntaxique de l'intonation: l'intensité joue un rôle secondaire (chute rapide par exemple de l'intensité en fin de phrase); mais la position des pauses, leur longueur, la durée de chaque syllabe de la phrase jouent également un rôle très important. Les rapports entre les facteurs prosodiques et la structure syntaxique des phrases prononcées sans emphase semblent apparaître plus nettement en français que dans les autres langues étrangères et ce fait est sans doute dû à la non-existence en français d'un "accent"(ou "stress") distinctif au niveau du mot: un enfant français peut lire,la plupart du temps, très correctement un mot qu'il n'a jamais entendu auparavant, ce qui n'est pas le cas pour la majorité des autres langues. La fréquence du fondamental semble, en français , libérée de contraintes lexicales au niveau du mot et son rôle purement syntaxique en est augmenté.

* Contribution à la synthèse par règles du français. Thèse de troisième cycle. 1971.

1) EXEMPLES DE PHRASES COURTES:

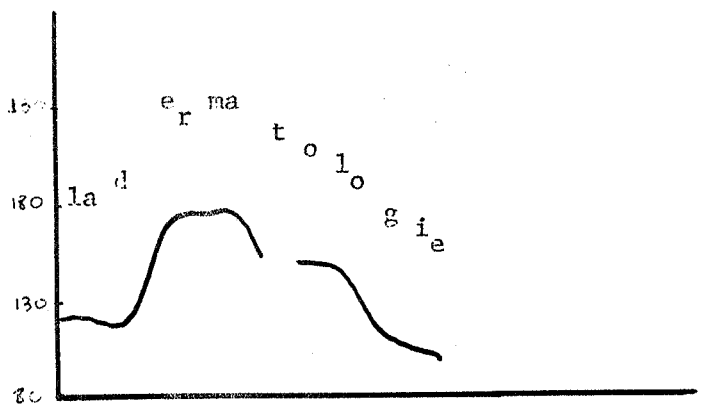


Je pars en vacances.

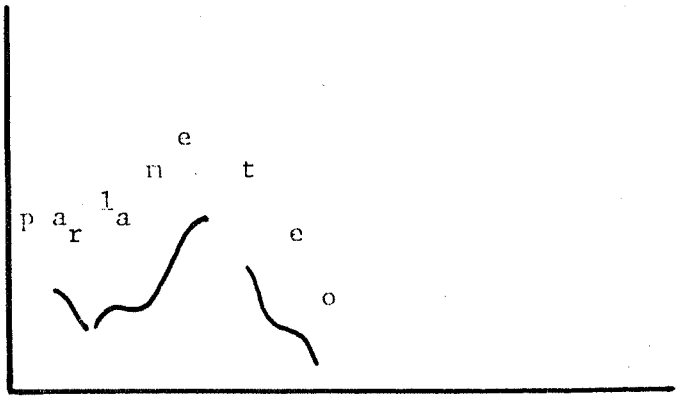


Il etudie la dermatologie.

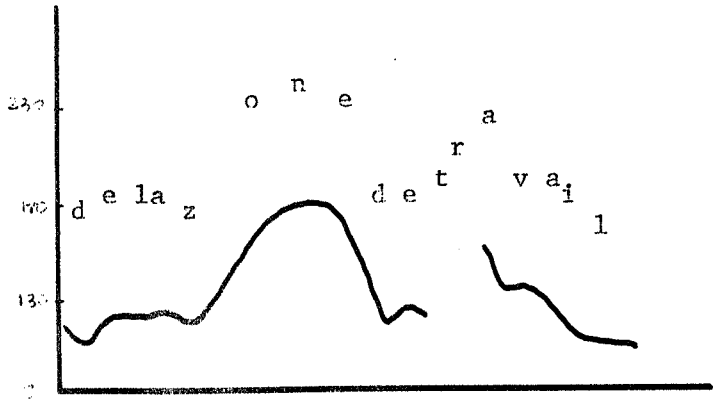
2) EXEMPLES DE FINS DE PHRASES:



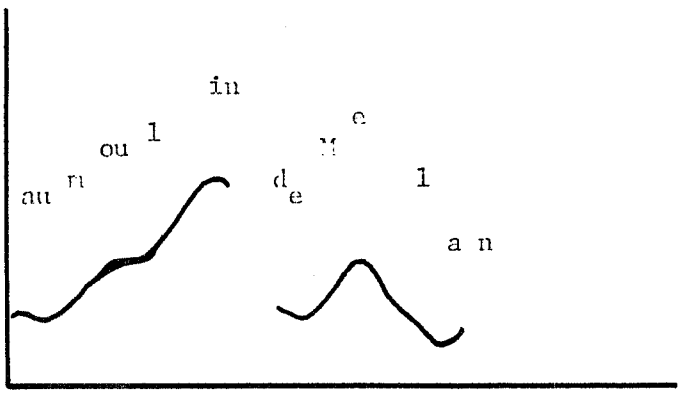
... la dermatologie.



... par la meteo.



... de la zone de travail.

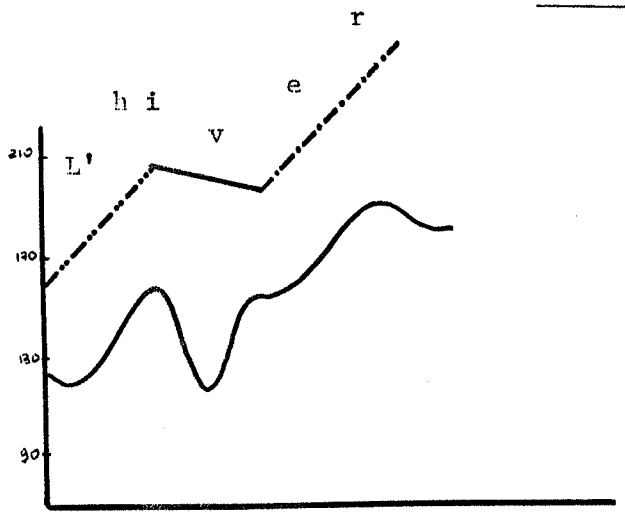


... au moulin de Melan.

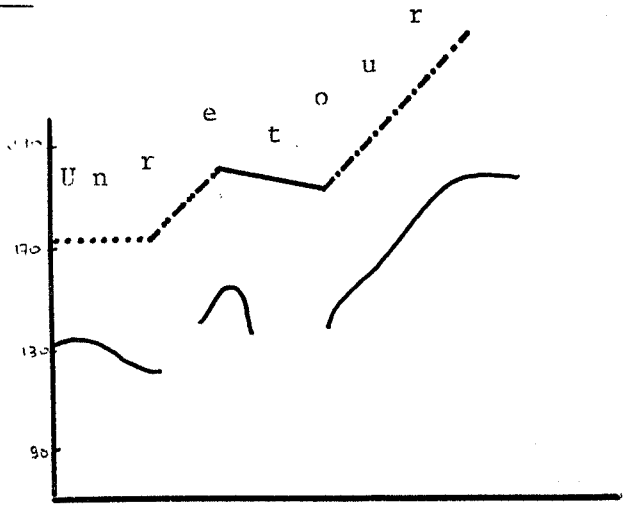
-FIGURE 1-

100 msec

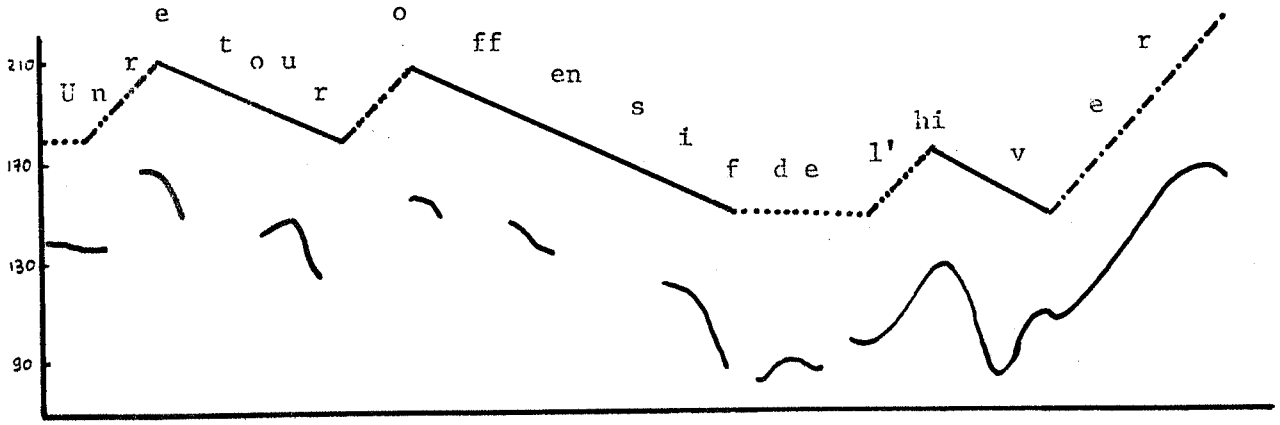
-FIGURE 2-



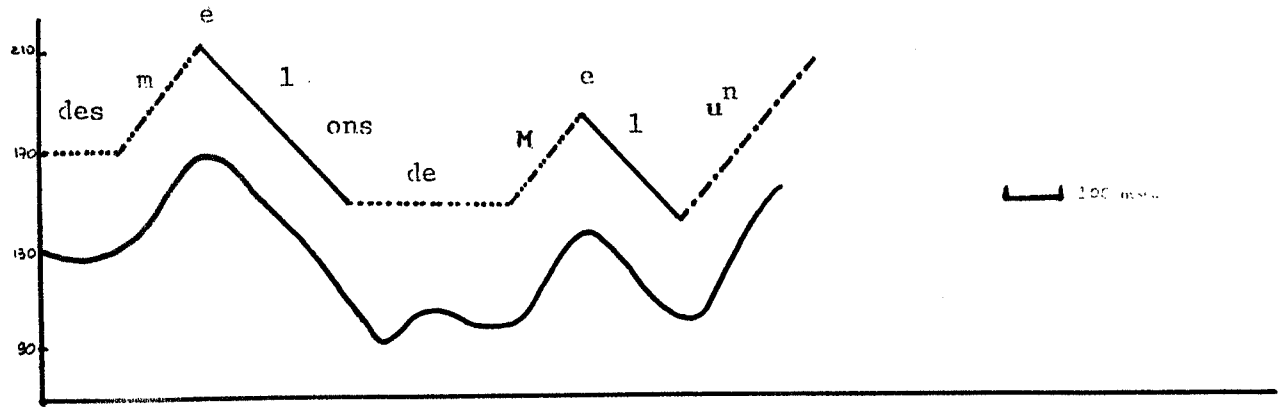
L'hiver (sujet)



Un retour (sujet)



Un retour offensif de l'hiver (groupe sujet)



... des melons de Melun (groupe objet)

Legende :

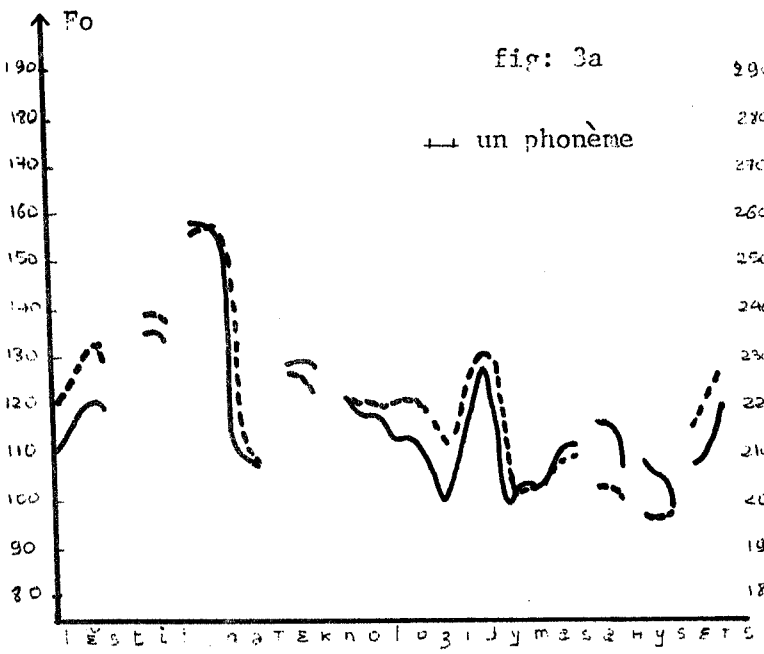
Impulsion de debut de mot lexical

Impulsion de la derniere syllabe du groupe

Pente de declinaison

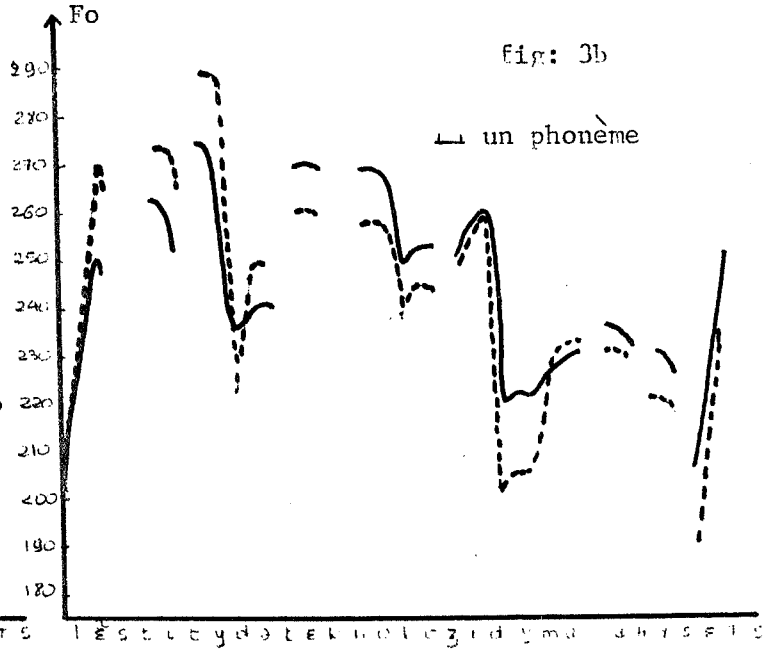
Mot fonctionnel

-FIGURE 3-



2 locuteurs (PD et PB), rythme "normal"
---: locuteur PD
——: locuteur PB

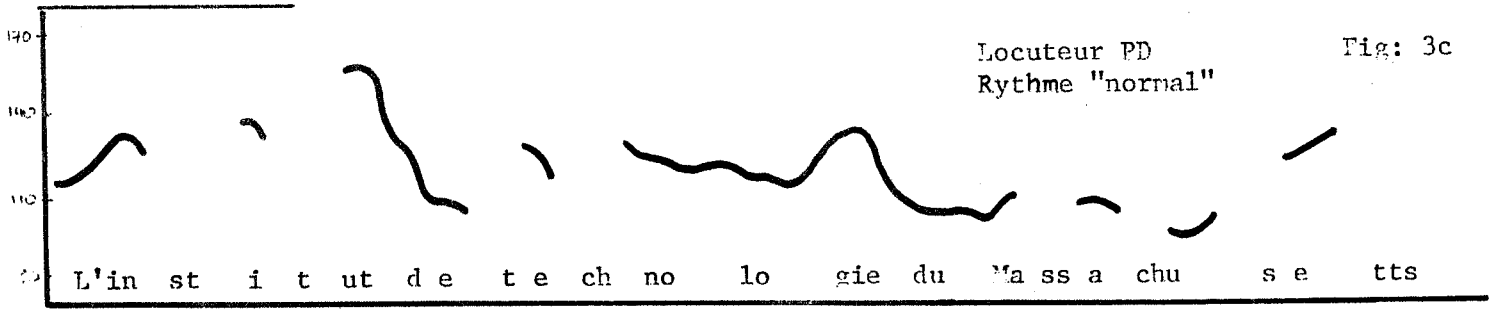
Voix masculines

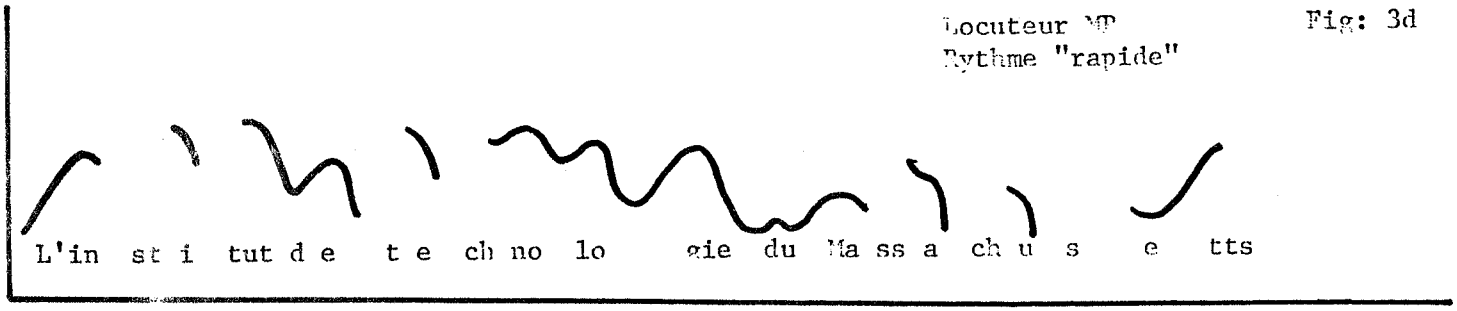


Même locuteur (MF):
---: rythme "rapide"
——: rythme "normal"

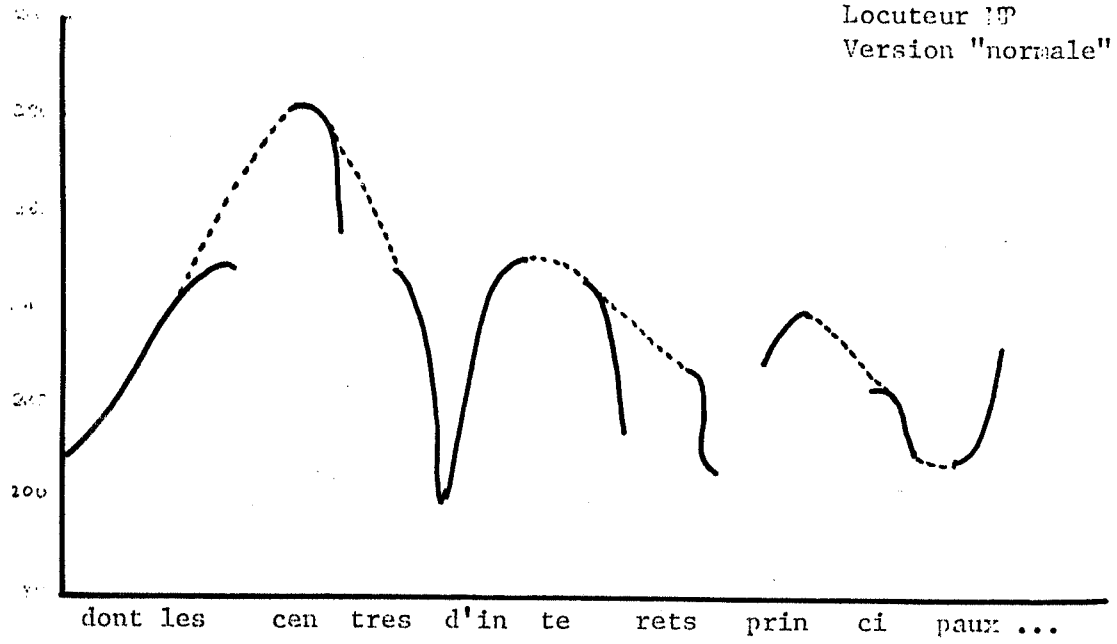
Voix féminine

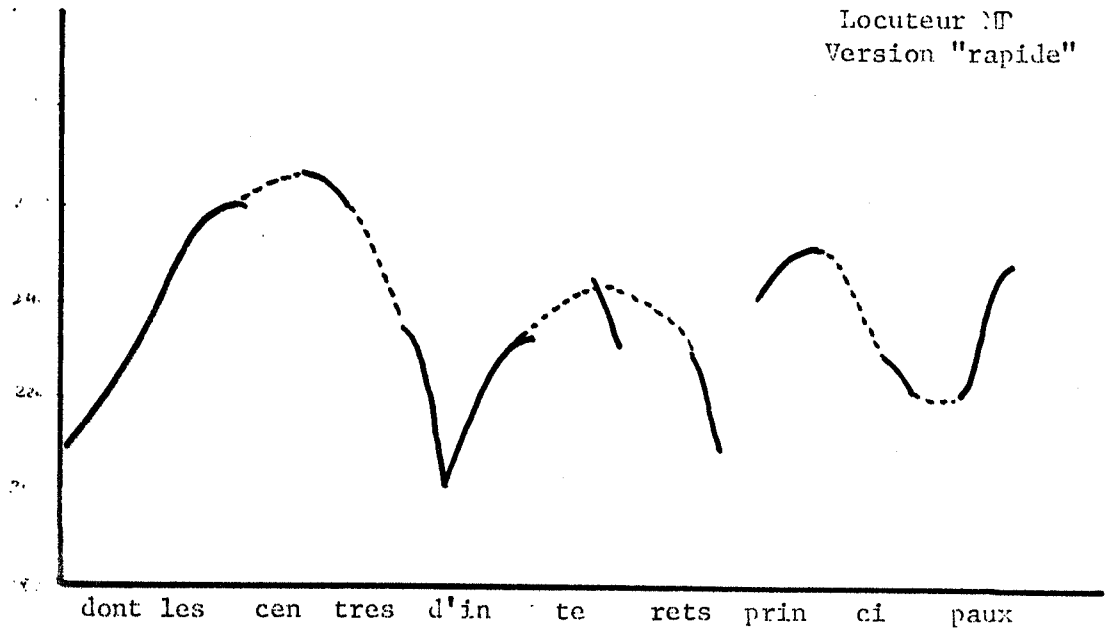
E X E M P L E S :





-FIGURE 4-





une syllabe

SYNTHESE PARAMETRIQUE DE L'INTONATION DE LA PHRASE ENONCIATIVE EN FRANÇAIS

Danièle LARREUR
C.N.E.T.
Lannion

et

Louis-Jean BOE
Institut de Phonétique
Grenoble

Résumé

Pour le français, on peut considérer comme satisfaisante la procédure d'analyse prosodique qui a pour but l'élaboration de schémas intonatifs fréquentiels positionnés par rapport à un certain nombre de niveaux. Cette méthode permet de rendre compte simultanément de deux systèmes aux fonctions complémentaires, et de remédier par ailleurs à certaines de leurs lacunes. Dans une étude précédente ont été spécifiés, pour la phrase énonciative, trois niveaux, définis par rapport à la fréquence laryngienne moyenne, comme des zones de fréquence dans lesquelles l'attaque, le maximum et la fin du contour intonatif ont une certaine probabilité d'occurrence. Ces résultats ont été utilisés et testés à partir d'un vocoder dans le cadre de la synthèse paramétrique du français, les variations de la fréquence laryngienne étant considérées comme le résultat de contributions relatives à l'intonation de la phrase et aux caractéristiques intrinsèques.

Summary

As far as French is concerned, the process of prosodic analysis, for aiming at the elaboration of intonational patterns related to a certain number of pitch levels, can be considered as satisfactory. This approach uses the advantages of two complementary systems and provides solutions for two other lacks. In a previous paper three pitch levels have been specified for statement in relation to the mean laryngeal frequency as frequency bands in which the attack, maximum and the end of the intonational pattern have an occurrence probability. These results have been used and tested with a channel vocoder in the frame of the parametric synthesis of french ; the variations of laryngeal frequency have been considered as the result of the contributions of sentence intonation contour and intrinsic characteristics.

Texte complet : voir le volume II des Actes

UNE METHODE DE SYNTHESE PAR REGLES DU SIGNAL VOCAL DANS SA REPRESENTATION AMPLITUDE-TEMPS

Xavier RODET

COMMISSARIAT A L'ENERGIE ATOMIQUE
CENTRE D'ETUDES NUCLEAIRES DE SACLAY
SERVICES D'ELECTRONIQUE DE SACLAY
Service d'Electronique pour la Recherche et les Applications
Section d'Assistance Electronique à la Recherche Fondamentale

B.P. n° 2 - 91190 GIF SUR YVETTE

RESUME

Cet article décrit une méthode de synthèse par ordinateur du signal vocal à partir d'informations sur les phonèmes, codées dans un minimum de place en mémoire. Les zones stables du signal sont obtenues par répétition périodique (voyelles et consonnes voisées) ou aléatoire (fricatives sourdes) d'une très courte séquence enregistrée. Les transitions sont calculées suivant une nouvelle méthode en fonction de la voyelle associée.

A METHOD FOR THE SYNTHESIS BY RULE OF VOCAL SIGNAL IN IT'S AMPLITUDE-TIME REPRESENTATION

ABSTRACT

The present paper deals with a research on computer synthesis of the vocal signal. Informations on phonemes are coded and request a minimum storage. Stable parts of the signal are produced by periodic or random repetition (voiced sounds or unvoiced fricatives) of a very short recorded sequence. A new method is proposed to compute transitions as functions of the joined vowel.

1 - INTRODUCTION

a) But de l'étude

Nous cherchons à réaliser une unité à réponse vocale, simple, implantable sur un mini-ordinateur et ne nécessitant qu'un appareillage analogique réduit. Nous entendons par "unité à réponse vocale" un système transformant en parole une suite codée de phonèmes.

b) Traitement du signal (cf./4/)

Nous travaillons sur le signal acoustique lui-même fonction du temps. Son amplitude est échantillonnée à 10 kHz et sur 15 niveaux seulement. La figure 1 montre les informations Ai et Δt_i conservées par le codeur. Δt_i est le nombre d'échantillons entre deux extremum du signal et A_i est l'amplitude de l'extremum ($-7 \leqslant a_i \leqslant +7$).

L'écoute du signal obtenu à l'aide de ces seules informations se montre satisfaisante sur le plan de la compréhension quoique la qualité souffre du traitement effectué. Mais la compression de la quantité d'informations qui en résulte est de l'ordre de 2 par rapport au signal échantillonné à 10 kHz sur 15 niveaux. Cette qualité représente pour nous une référence à approcher en synthèse. Actuellement, les niveaux d'amplitude sont régulièrement espacés ; nous nous proposons de chercher si une autre répartition, par exemple logarithmique, ne donnerait pas de meilleurs résultats.

c) Moyens utilisés

Nous utilisons un appareillage simple, relié à un ordinateur MULTI 20 (figure 2). En entrée, le signal provenant du microphone est codé sous forme de temps entre les passages par zéro de la dérivée notés en dix millième de seconde sur quatre bits, et d'amplitudes maximales entre ces passages par zéro, sur quatre bits également.

En sortie, à partir des informations successives de longueur et d'amplitude, le signal est restitué suivant une simple interpolation linéaire, suffisante pour une bonne compréhension. Les courbes représentant le signal acoustique en fonction du temps sont visualisées sur une console graphique.

2 - PLUSIEURS METHODES POSSIBLES DE SYNTHESE

a) Enregistrement de messages

La compression de l'information obtenue (applicable à tout autre signal) permet d'envisager l'enregistrement de messages, sur un disque magnétique par exemple, et leur restitution à volonté au moyen d'un mini-ordinateur et du dispositif de sortie indiqué, extrèmement simple. Il faut alors 60 K octets environ par minute de parole enregistrée (débit d'information : 8.000 bits/seconde).

b) Synthèse par diphonèmes

L'examen sur console graphique des courbes sonores enregistrées permet de segmenter les diphonèmes en des endroits précis, tels que leur juxtaposition ultérieure n'altère pas le pitch.

On peut alors évaluer à 150 K octets la capacité de stockage nécessaire.

c) Juxtaposition des phonèmes : ses limites

Il est bien connu que la synthèse est impossible par simple juxtaposition de phonèmes, nombre d'entre eux (plosives par exemple) n'ayant aucune existence autonome. Cependant, la segmentation de certains phonèmes en des endroits précis, respectant la régularité du pitch et où le signal s'annule, permet de les recoller de telle sorte que l'écoute soit naturelle et sans bruit parasite ; la compréhension de la consonne est encore possible.

Par exemple, la courbe d'un "M" extraite d'une syllabe M-voyelle peut être recollée à celle de n'importe quelle voyelle enregistrée isolément. A l'écoute, la syllabe obtenue est reconnaissable pour une oreille entrainée.

3 - SYNTHESE PAR REGLES

a) Mémorisation et synthèse des voyelles et consonnes voisées

Dans la partie moyenne des phonèmes voisés, la fréquence d'excitation des cordes vocales et la fonction de transfert du conduit vocal varient peu. Ainsi, la courbe du signal sonore résultant est quasi-périodique à la fréquence du pitch. Elle se présente comme la répétition à des intervalles de temps presque constants d'un motif presque invariant (figure 3).

Nous avons donc synthétisé les parties stables des phonèmes voisés par répétition d'une pseudo-période prise entre deux pitch d'un phonème enregistré. Pour les voyelles, nous avons du tenir compte d'une enveloppe définissant l'amplitude moyenne, croissante, stable puis décroissante en première approximation (figure 4). La voyelle ainsi produite est compréhensible mais souffre de l'invariance des paramètres (en particulier absence de fréquences inférieures à celle du pitch). Plusieurs améliorations ont été apportées :

- faibles variations de la fréquence du pitch au cours de l'évolution de la voyelle (moins de 10 %)
- faibles variations de l'enveloppe (sinusoïdales par exemple, à une fréquence basse)
- superposition d'une basse fréquence (inférieure à celle du pitch).

Ainsi, la mémorisation d'une voyelle est ramenée à celle d'une seule période (notée sur une vingtaine d'octets en moyenne), ce qui constitue une compression de l'information d'un facteur supérieur à 10 par rapport à la mémorisation d'une voyelle complète. On a ainsi obtenu toutes les voyelles avec une bonne qualité compte tenu des limites du système.

Si l'intervalle de répétition choisi est plus court que l'intervalle origine, le motif est tronqué en conséquence.

Dans le cas contraire, le motif est complété par une portion de courbe d'amplitude nulle (figure 5). Toute variation de la fréquence du pitch (mélodie) est donc possible. On sait que les voix de femmes se différencient essentiellement des voix d'hommes par la fréquence (double environ) du pitch, tandis que les fréquences des formants varient peu (17 % environ). Nous avons pu, à partir d'une même séquence, produire indifféremment des voyelles correspondant à des voix d'hommes ou de femmes.

b) Fricatives sourdes

Dans le cas de ces consonnes non voisées (s, ch, f) on peut considérer que la source d'excitation du conduit vocal est produite par le frottement de l'air dans un rétrécissement de ce conduit. Il en résulte un spectre continu dans une certaine bande de fréquences, caractéristique de la consonne. Nous avons donc synthétisé la partie bruitée de ces consonnes de la façon suivante. Une portion seulement de l'enregistrement de la consonne est gardé en mémoire. Elle est découpée en plusieurs segments délimités par des points ou le signal s'annule. La consonne est alors synthétisée par juxtaposition de ces segments dans un ordre aléatoire.

c) Transitions consonne-voyelle

Nous exposons ici une méthode de synthèse des consonnes en cours d'étude dans notre laboratoire. Elle consiste à calculer la transition du début ou de la fin de la voyelle, caractéristique de la consonne. Notons que la transition vient en complément des parties voisées ou bruitées pour les consonnes voisées ou les fricatives sourdes.

La prononciation des consonnes met en jeu une déformation du conduit vocal, donc une modification de sa fonction de transfert. Ce qui se traduit, en termes de fréquences, par des variations formantiques /2/, /3/. Nous observons, sur la courbe du signal vocal, des variations de fréquence du pitch, et de forme du motif de la voyelle. Nous cherchons à reproduire ces variations de la façon suivante :

- le motif caractéristique d'une voyelle est décomposé en deux courbes (figure 6). L'une C_1 est obtenue par lissage de la courbe origine C_0 l'autre par différence, point par point, $C_{2_i} = C_{0_i} - C_{1_i}$. En première approximation C_1 correspond au premier formant et C_2 au second et troisième /5/, /6/,

- ces deux courbes peuvent subir séparemment des anamorphoses (figure 7 a) suivant l'axe des temps, ce qui revient à augmenter ou diminuer la fréquence de l'un ou l'autre formant*. Les amplitudes peuvent être également multipliées par des coefficients variables tenant compte de l'instant et de la rapidité d'apparition des formants /1/. C'est la somme de deux courbes après traitement qui constitue alors une période entre deux pitchs, (figure 7b). Ainsi, de période en période, des déformations sont apportées au motif origine de la voyelle, traduisant la transition consonne-voyelle.

---

* Remarque : Nous avons conservé le terme de "formants", mais il s'agit précisemment de fréquences associées au signal entre deux pitchs.

- Les variations de pitch sont traitées comme au § 3 -a).

Les calculs devant être effectués en temps réel par un mini-ordinateur, nous essayons, ce qui peut être discuté, de caractériser les consonnes par des règles de variations des paramètres identiques quelque soit la voyelle jointe. En plus,de ces règles de base, quelques améliorations peuvent être apportées, comme la superposition d'une oscillation de basse fréquence traduisant l'ouverture d'une occlusion du canal vocal, ou un certain déphasage des courbes C_1 et C_2 , qui doit intervenir dans les transitions /7/. Nous avons déjà obtenu, en simulation, quelques consonnes (p, k, m, s).

CONCLUSION

Nous cherchons à définir toutes les transformations d'un motif de voyelle nécessaires à la synthèse des consonnes. Ceci afin d'établir sur un mini-ordinateur le système qui calculera en temps réel toute succession de phonèmes rendant compte de la parole.

REMERCIEMENTS

Nous remercions le Professeur J.C. SIMON de l'Université PARIS VI, Conseiller Scientifique aux Services d'Electronique de Saclay, qui nous a guidés dans ces recherches, ainsi que Monsieur de COSNAC des Services d'Electronique de Saclay qui a dirigé cette étude.

Nous remercions également Messieurs M. BAUDRY et B. DUPEYRAT pour leur fructueuse collaboration.

B I B L I O G R A P H I E

/1/ ALINAT - "Reconnaissance de phonèmes en temps réel en vue de réaliser des liaisons à faible débit - Application à la sténotypie automatique" Rapport final d'étude DRME n° 198/71 - 1971

/2/ L. SANTERRE - "Transitions articulatoires et transitions acoustiques dans la parole réelle"
Compte rendu des Journées d'Etude sur la Parole
31 Mai - 2 Juin 1972 au CNET LANNION

/3/ G.I. TSEMEL - "Application in speech recognition of some data on anditory segmentation and speech wave parameters' perception
Proc. of the symp. Auditory Analysis and Speech Perception
August 1973 - LENINGRAD

/4/ M. BAUDRY, B. DUPEYRAT, C. FRANK - "Reconnaissance automatique de la parole, étude de la segmentation"
Compte rendu des Journées d'Etudes sur la parole
31 Mai - 2 Juin 1972 au CNET LANNION

/5/ R.W.A. SCARR - "Zero crossings as a means of obtaining spectral information in speech analysis"
I.E.E.E. Transactions on audio and Electro-acoustics Vol. AU-16 n°2
June 1968

/6/ MABO ROBERT ITO, R.N. DONALDSON - "Zero crossing Measurements for analysis and recognition of speech sounds"
I.E.E.E. Transactions on audio and Electro-acoustics VOL. AU-19 n° 3
September 1971

/7/ J. BOSQUET - "Perception auditive et analyse spectrale"
Actes des 4ème Journées d'Etudè du Groupe de la "communication parlée"
BRUXELLES Mai 1973

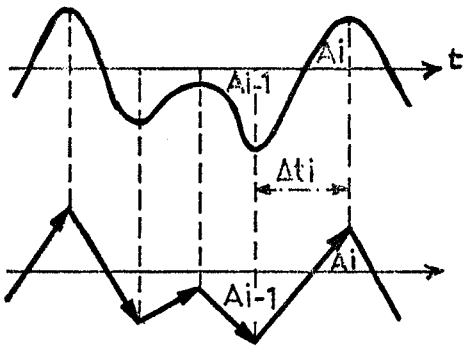


Fig 1

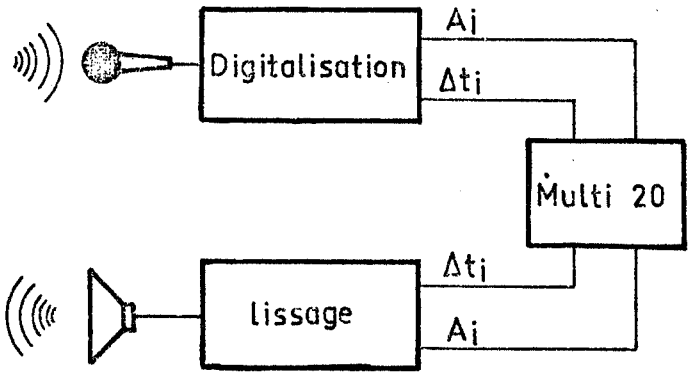
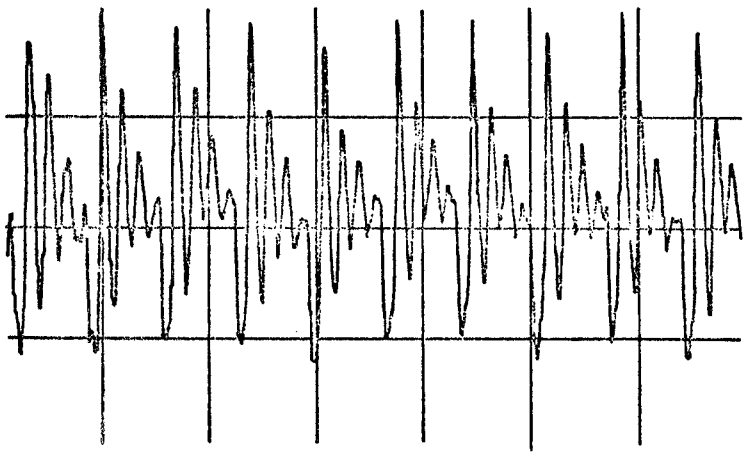


Fig 2



Courbe amplitude temps pour la partie stable d'une voyelle (/o/ de Lo)

Fig 3

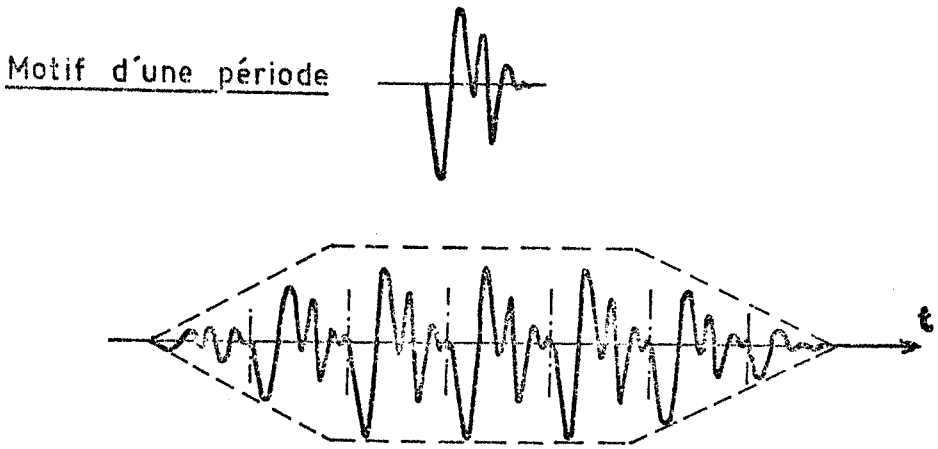


Fig 4

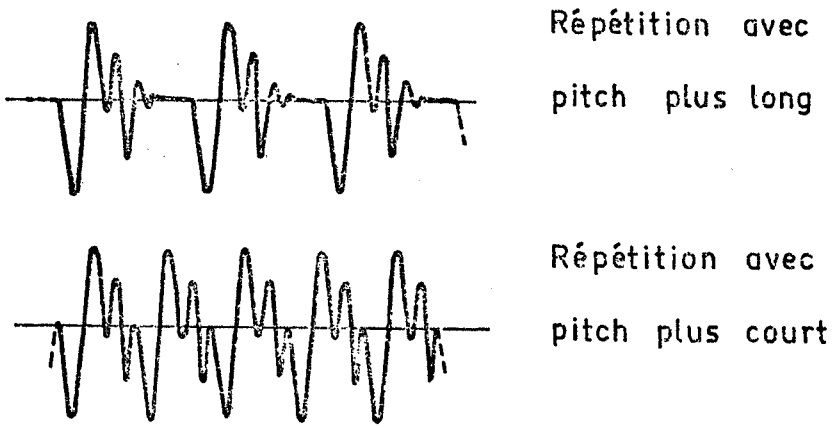


Fig 5

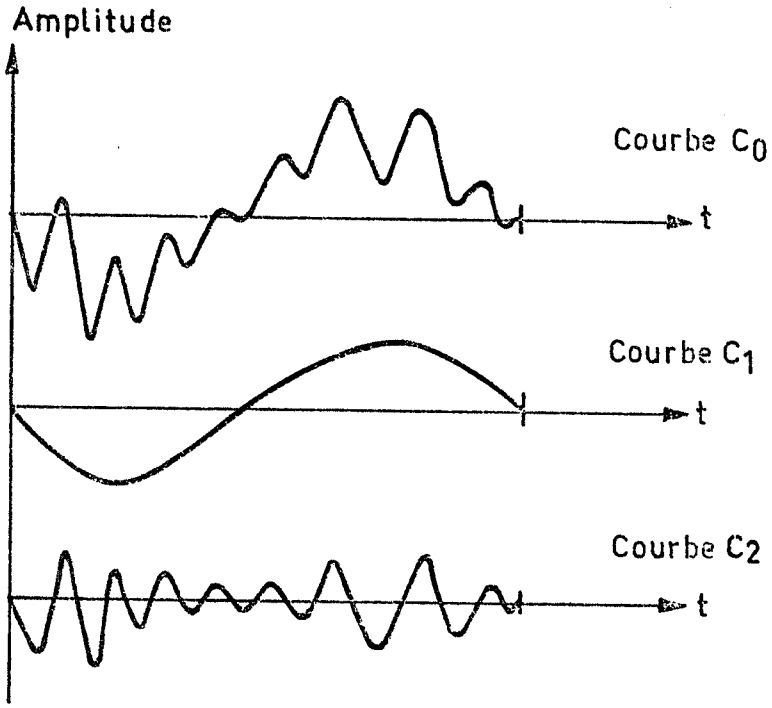
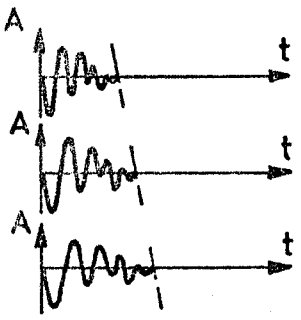


Fig 6



Courbe anamorphosée $C'(t) = C(k'.t)$ $k' > 1$

Courbe origine $C(t)$

Courbe anamorphosée $C''(t) = C(k''.t)$ $k'' < 1$

Fig 7a

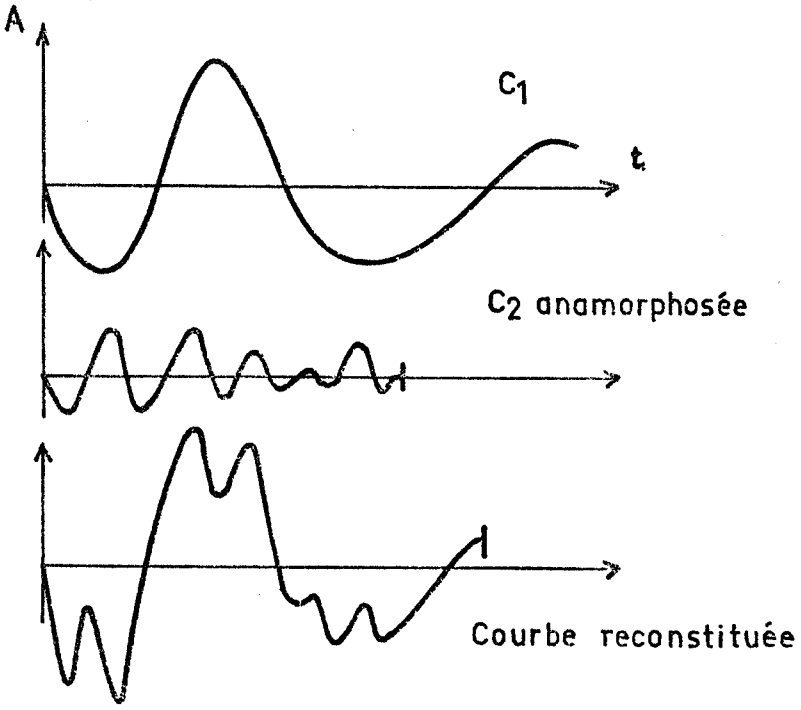


Fig 7b

A.R. Meo, M.Mezzalama, E. Rusconi.

A SPEECH SYNTHESIS SYSTEM

Centro di Studio per l'Elaborazione Numerale dei Segnali (CNR)

Politecnico di Torino

Istituto elettrotecnico Nazionale "G. Ferraris"

A SPEECH SYNTHESIS SYSTEM

A.R. Meo, M.Mezzalama, E. Rusconi.
CENS/IENG/IEG Politecnico di Torino
IOI29 TORINO - Italy.

Résumé

On décrit notre projet de synthèse par règles. Le système de synthèse utilise un calculateur numérique pour traduire le texte écrit en un ensemble de signaux de commande actionnant un synthétiseur à formants. Celui-ci est réalisé au moyen d'un calculateur spécialisé.

Les paramètres de chaque phonème sont les coefficients d'un filtre digital, exprimé comme une fraction rationnelle dans le plan z.

Les lois de transition sont mémorisées dans une matrice et sont exprimées sous forme de courbes normalisées, qui ne sont pas nécessairement des fonctions mathématiques.

Les valeurs des paramètres peuvent être modifiés par l'introduction de règles allophoniques.

On a prévu le contrôle des caractéristiques suprasegmentales, qui sont à l'étude dans une recherche parralèle sur la voix naturelle.

Abstract

A general approach to speech synthesis is described.

A special-purpose computer, operating as a resonance synthesizer, provides the continuous real-time ontput of speech, while a general-purpose computer translates a written text into a suitable set of commands, deduced from stored rules, for the synthesizer.

The phoneme parameters are stored as the z-transform coefficients of the linear system simulated by the resonance synthesizer and a matrix of transition tables gives the motion law for any pair of phoneme parameters.

The suprasegmental features are controlled and some intonation patterns may be introduced.

These are object of a parallel research in which duration, amplitude and pitch are modified and tested in time domain, using natural speech.

In this communication our speech synthesis project is described.

Although our aim is to produce synthetic Italian speech and a great amount of work has been done in order to extract the parameters of Italian by means of a speech analysis digital system [Ref. 1], our synthesis by rule system is designed as general as possible, so that it could run for any language provided that suitable parameters are introduced.

The system works on a g.p. computer DDP-516 which processes the rules and computes the parameters. These may be employed for synthesizing speech by the DDP-516 it-self, or may be communicated to a special purpose computer, specifically built for this purpose, (S.P.C.) [Ref.2] which performs speech synthesis in real time operating as a digital filter whose configuration may be programmed.

SYNTHESIS STRATEGY

Referring to some leading works on the field, Rudnicki's at M.I.T. [Ref.3] and Mattingly's at Haskins Lab [Ref.4], we designed the synthesizer both as a cascade and a parallel system of second-order resonant filters.

The synthesizer, in both the software and hardware versions, works as a digital system, with an output digital-to-analog converter and an audio-amplifier to produce the speech-like sounds. In this way the rules employed will have a universal value, since they are not related to a particular analog terminal.

The synthesis is performed by means of the following steps.

a) The ordinary (Italian) printed test (OPT) is converted into a phonetic element sequence (PES) and at the same time the intonation pattern is introduced inserting stress information (pitch and amplitude values) into the phonetic sequence. Generative rules able to control prosodic features are not yet operating in the synthesis system because a theory of intonation is not available for the Italian language. Some partial results of the study of this problem will be described in a next section.

Since relatively unambiguons pronounce rules exist for Italian, the translation from OPT to PES is made automatically, while few exceptions and sound differences, mainly related to regional characteristics, may be easily overcome or justifiably ignored.

b) The parameters associated to each phoneme are modified by mean of allophonic rules. Lengthening or shortening of a vowel duration depending on the following phoneme is an example of such rules. Although this step is logically the second one in a synthesis system most of the rules to be used can be investigated only when the whole system is already running.

c) The control signals (CS) for the digital synthesizer are computed starting from the PES. This step is the central one and will be described with more details.

The rules are contained in a set of look-up tables:

c1) Phoneme parameter tables (PPT). In phonemic tables at last two different kinds of parameters may be stored; specifically, either the classical acoustic parameters (i.e. formant frequencies, band widths and gains) or the z-transform coefficients

of the system transfer function may be used. At the moment we prefer to use the second kind of parameters because a more direct control of the system response is possible and a save of computation time is attained in the synthesis process.

In our system, 70 phoneme tables have been used. Each of these contains the phoneme name, its duration and amplitude, and the excitation characteristics, in addition to the z-transfer coefficients.

The system may be excited by periodic pulses or by noise.

Same glottal pulse waveforms (G P W), for different voiced sounds, will be available and are stored in tables.

A last phoneme parameter allows a combination of noise and G P W.

c2) Transition Tables (T R T). Several tridimensional matrices give the motion rules for each control parameter of phoneme couples or of couples of phoneme sets.

A transition rule is specified by the motion duration, by the starting point referred to the end of the current phoneme and by the motion curve given as a normalized ten-point curve. Also the normalized motion curves are in look-up tables.

In fig. 1 an example of a transition computation is shown.

The control parameters computed in such a way are stored in a mass memory and then provided

to the synthesizer that can produce speech in real time (using a hardware terminal).

SUPRASEGMENTAL FEATURES

The study of the rules for suprasegmental features (SF) is based on the results obtained and the tests made in a parallel research project on speech synthesis by segments.

We decided to work in this direction taking into account the strong syllabic structure of Italian /Ref.5/ and our efforts were made to overcome the main limit of this approach to artificial speech production, i.e. the difficulties to modify the S.F. of the synthetic speech. In our work a set of modifications are performed in the time domain, starting from natural speech.

The suprasegmental features used as prosodic elements include intensity, duration and fundamental period.

Dealing with digital-samples signals, the modification of intensity and duration presents no problem. The modification of pitch is obtained as follows: every natural period is artificially lengthened, inserting a segment of a suitable signal (e.g.: simply balnks). The effect of modifying the pitch is easily imaginable, but we have to consider how the whole signal is affected by this manipulation.

The tests made in the Italian language show that to the ear the new signal appears a bit un-natural, but is completely intelligible, and the supra-segmental modifications are fairly perceptible.

The method results fairly useful for the study of

the intonation rules thanks to the ease with which the experimenter can perform tests having different prosodicfeatures. The preparatory work consists only in the selection of the instants at which the blank signal is to be inserted. They must be chosen so that the difference between a natural lengthening of the pitch and our artificial lengthening may be interpreted as a quantization error affecting the less intense parts of the waveform.

REFERENCES

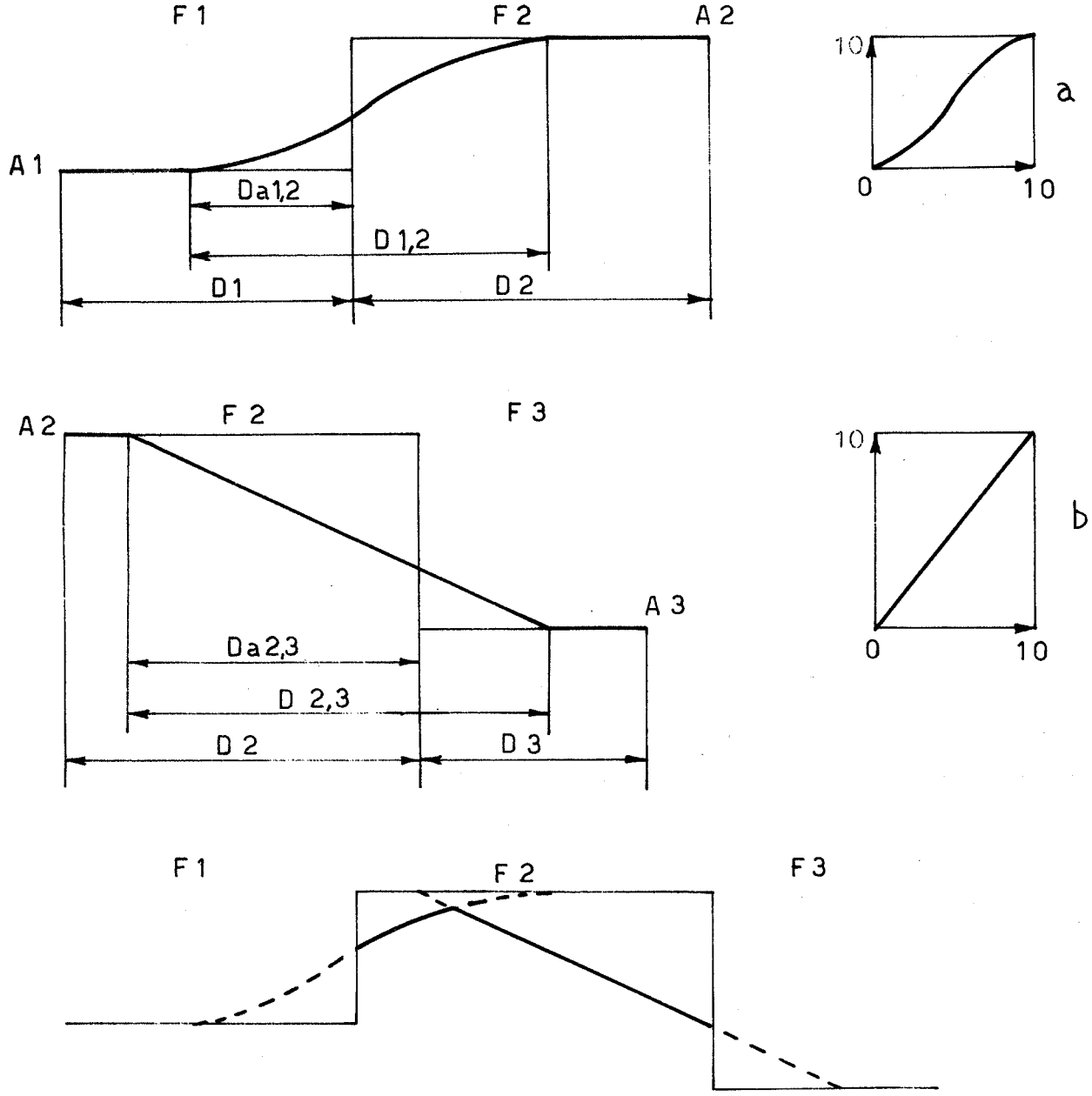
(1) M.Guglielmo, A.R.Meo, M.Mezzalama:"C.V.S.: a technique for generating a type of visible speech based on synchronous spectral analysis" (44th Audio Eng.Soc.Convention,Rotterdam, February 1973).

(2) R.De Mori, S.Rivoira, A. Serra: "A programmable bank of digital filters for high-frequency signal processing" (Proc.Seminar on Digital Filtering, Florence, September 1972).

(3) L.R. Rabiner: "A model for synthesizing speech by rules " (IEEE, Trans. on Audio 1969).

(4) J.G. Mattingly:"Synthesis by rules of G.A. English"(Status Report on Speech Research, Haskins Laboratories, April 1968).

(5) Francini G.L., Debiasi G.B.,Spinabelli R.D.:"Study of a System of Minimal Speech - Reproducing Units for Italian Language" - JASA 43-6 June 1968.



F1.F2,F3 input P E S.
a, b transition normalised curves.
A_i steady state of the parameter a_n.
D^i duration of the phoneme F_i.
$D^i_{j,j}$ duration of transition between F_i and F_j.
$Da^j_{i,j}$ starting coordinate of the transition $F^j_i - F_j$.

Fig. 1 - Motion of the parameter a_n for the phoneme F_2.

L'UNITE A REPONSE VOCALE ICOPHONE V

par D. TEIL *, M. CASTELLENGO **, J SAPALY **

* Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur, C.N.R.S., - Orsay

** Laboratoires d'Acoustique et d'Electronique, Institut de Mécanique Théorique et Appliquee, - Université Paris VI

Résumé

Dans le cadre de nos études sur la synthèse de la parole, nous avons conçu et réalise un terminal à réponse vocale de dimensions reduites, au fonctionnement simple et à vocabulaire illimité.

Le message à synthétiser est donné sous forme orthographique. Il est transformé en une suite phonétique qui permet l'assemblage des diphonèmes mémorisés sous forme binaire.

Les règles de synthèse sont implicitement formulées dans la forme temps-fréquence des diphonèmes.

La synthèse est quasi-instantanée. L'appareil peut être relié à n'importe quel ordinateur muni d'une prise télétype.

Il peut également fonctionner de manière autonome.

THE VOCAL REPONSE UNIT ICOPHONE V

Summary

Our speech synthesis studies lead us to design and build a vocal reponse unit. With its small dimensions and its unlimited vocabulary, it is of particularly easy use.

The message to be synthesized is given in orthographic form. It is transformed into phonetic codes which permit to link the memorized diphones.

The synthesis rules are given by the time-frequency pattern of the diphones.

The synthesis is immediate. The device can be connected to any computer which has a teletype input-output. It can also work off-line.

L'UNITE A REPONSE VOCALE ICOPHONE V

par D. TEIL *, M. CASTELLENGO **, J SAPALY **

* Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingenieur, C.N.R.S., - Orsay

** Laboratoires d'Acoustique et d'Electronique, Institut de Mécanique Théorique et Appliquee, - Université Paris VI

I La Synthèse par Diphonèmes

Des recherches commencees en 1966 ont conduit à la realisation de deux générateurs de parole synthetique à commande optique: Les ICOPHONES I et II, qui ont permis de verifier nos hypothèses sur l'utilisation des diphonèmes en synthèse de parole.(1).

La synthèse en temps réel a ensuite été abordée à l'aide de l'ICOPHONE III à commande numerique, couplé à un ordinateur IBM 1130 (2).

Cet ensemble nous a permis de terminer la mise au point du dictionnaire des diphonèmes de la langue française et de résoudre les problèmes informatiques (software et hardware) relatifs à la traduction automatique d'un message écrit en message parlé (3),(4).

Nous avons ensuite réalisé l'ICOPHONE IV qui, couplé à l'IBM 1130, nous a permis d'améliorer la qualité de certains éléments phonétiques par l'utilisation de générateurs de bande de bruits et de commencer une étude sur le timbre et l'intonation.

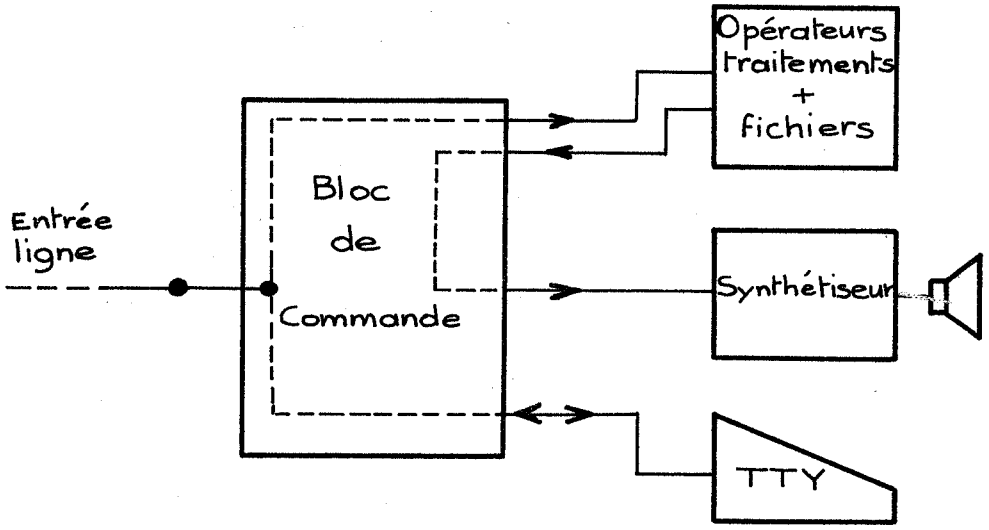
II l'unité a reponse vocale

L'appareillage actuel du L.I.M.S.I.* constitue un puissant outil de recherche en synthèse de la parole.

Parallèlement à la recherche fondamentale qui se poursuit avec cet appareillage, nous avons mis au point, avec l'aide de la Délégation à l'Informatique, une unité à réponse vocale exploitable industriellement: L'ICOPHONE V. Cet appareil transforme en parole parfaitement intelligible un texte fourni sous la forme alphabétique habituelle. Le vocabulaire est illimité, la réponse immédiate. Le débit requis sur la ligne de commande est minime (de l'ordre de 50 bands), et l'appareil peut fonctionner soit de manière autonome en liaison avec un simple télétype, soit comme un terminal d'ordinateur. Dans ce dernier cas la connexion est effectuée au moyen de la sortie "télétype" de l'ordinateur.

III Description d'ensemble

L'Icophone V est constitué de 4 parties:



Le bloc de commande constitue la partie qui a pour rôle la gestion des différents modules. C'est l'unité centrale d'un mini-calculateur (ORDOPROCESSEUR), dont la microprogrammation a été réalisée en collaboration avec TITN.

Les opérateurs de traitement assurent les fonctions suivantes:

- Transformation du message alphabétique en une suite de symboles phonétiques codés,

- Reconstitution du squelette informatif du message par assemblage de diphonèmes schématisés extraits du dictionnaire,

- Edition de cette configuration binaire sur les oscillateurs de l'ICOPHONE,

Le synthétiseur est une réplique simplifiée de l'ICOPHONE IV. La fréquence du fondamentale est fixe (100 Hz); la vitesse d'élocution et l'amplitude globale du message sont réglables manuellement.

Une machine à écrire du genre télétype assure le dialogue avec l'ordinateur, ou avec le synthétiseur en fonctionnement autonome.

IV Réduction du dictionnaire

Le dictionnaire de départ ayant servi à la synthèse de l'ICOPHONE III est composé de 627 diphonèmes divisés dans le temps en 20 évènements. (4). Le rangement en mémoire de ce lexique complet utilise 37620 mots de 16 bits.

La nécessité de faire fonctionner l'ICOPHONE V en temps reel nous a conduit à utiliser des mémoires mortes. Pour minimiser le coût de l'appareillage, nous avons cherché à réduire au minimum l'encombrement du dictionnaire.

Les premiers essais de réduction ont porté sur la discrétisation du temps. Les éléments phonétiques ont été découpés selon divers pas de temps compris entre 4 et 40 ms par évènement. Les tests d'intelligibilité ont montré que l'optimum se situait aux environs de 10 ms. Notre choix s'est donc porté sur une discrétisation de 8 évènements par diphonème ce qui correspond sensiblement à cet optimum.

Les essais suivants ont consisté à regrouper des commandes d'oscillateurs pour réduire l'information fréquentielle. Pour que l'opération soit rentable il fallait passer de 46 à 32 commandes. Un tel résultat ne pouvait être obtenu qu'en changeant la distribution fréquentielle des oscillateurs ce qui conduirait à refaire une étude complète des diphonèmes et du synthétiseur.

Dans l'étape suivante, nous avons classé les diphonèmes en fonction de leur propriété de reversibilité pour ne ranger en mémoire qu'une transition sur deux. Certains diphonèmes sont entièrement reversibles, c'est le cas des éléments voyelle - voyelle et des éléments consonne - voyelle où la consonne est /m/, /n/, /l/, /j/.

Pour les autres éléments consonne-voyelle nous avons considéré que la transition vocalique est reversible ce qui a donné des demi-diphonèmes. Les consonnes sont reconstituées dans l'autre moitié. Pour simplifier nous utilisons les mêmes consonnes fricatives quelle que soit la voyelle. Les consonnes occlusives ont été classées en trois catégories suivant qu'elles sont associées à une voyelle qualifiée de grave (/o/, /ɔ/, /u/, /ɔ̃/), moyenne (/a/, /œ/, /ã/, /ɛ̃/), ou aigüe (/i/, /y/,/e/, /ɛ/). Dans les consonnes occlusives nous n'avons mis en mémoire que l'attaque définie sur l'évènement. Le "silence" qui précède (plus ou moins long, voisé ou non suivant la consonne) est introduit au moment de l'émission sur le synthétiseur.

Les éléments voyelle-consonne occlusive sont formés par la transition vocalique suivie d'un "silence" voisé ou non de 4 évènements.

Tous les évènements ont été comparés entre-eux et nous n'avons conservé en mémoire que les évènements différents.

Le lexique des formes phonétiques, composé d'un dictionnaire de phonèmes (12 voyelles, 3 fois 6 consonnes occlusives et 6 consonnes fricatives), d'un dictionnaire de diphonèmes entiers et d'un dictionnaire de demi diphonèmes, occupent 4111 mots de 16 bits dans l'élément mémoire de l'ICOPHONE V, et la table des identificateurs, 371 mots de 16 bits.

V Réalisation de l'ICOPHONE V

1 Les opérateurs de traitements et les fichiers:

Les fonctions à réaliser les plus importantes en volume sont la traduction phonétique, le rythme et l'adressage des formes phonetiques. La logique d'adressage assure la reconstitution des spectres à partir d'un algorithme de pointage dans une table et du lexique des formes phonétiques.

L'analyse de ces fonctions nous a conduit à utiliser un mini-ordinateur à programme figé du type ordoprocesseur qui répond parfaitement au problème:

- Par sa structure "bus" qui permet de connecter facilement toutes sortes d'opérateurs, en particulier l'opérateur "table" contenant le lexique des formes acoustiques, les opérateurs d'entrée (contrôleur ligne et contrôleur clavier) et de sortie (un ou plusieurs contrôleurs de synthétiseur).

- Par sa souplesse de code qui permet d'introduire des mini-instructions spécialisées en particulier pour l'algorithme de recherche dans la table.

- Par son coût économique qui permet d'obtenir un faible prix de revient.

2 Le Synthétiseur:

Il comprend sous forme de modules miniaturisés:

- 44 oscillateurs sinusoïdaux couvrant linéairement la gamme 100 Hz 4400 Hz,

- 3 générateurs de bande de bruit répartis dans la bande 1500 à 6000 Hz,

- 1 mélangeur

- 1 chaîne d'amplification et d'écoute.

Les oscillateurs fonctionnent en tout ou rien sur ordre de la logique de sortie. La fréquence et l'amplitude du signal de sortie sont réglés une fois pour toute, la constante de temps de la commande est ajustable en modifiant des composants discrets accessibles sur la carte, ce qui permet d'attenuer les fronts d'attaque des signaux. Les oscillateurs ne sont pas modulables en fréquence: on se limite à une voix voisée sans intonation. Les générateurs de bruit fonctionnent également en tout ou rien. On peut ajuster une fois pour toutes la fréquence centrale, la largeur de bande et l'amplitude du signal de sortie.

VI Les caractéristiques du terminal ICOPHONE V

Caractéristiques physiques:

- dimensions approximatives du terminal: 50 x 60 x 40 cm,
- organe de dialogue: machine à écrire,
- software intégré,
- reglages manuels de la vitesse d'élocution et de la puissance de sortie.

Caractéristiques de fonctionnement:

- autonome, en liaison avec un autre terminal du même type ou en liaison avec un ordinateur à la place d'un télétype,
- messages transmis en clair en français ou en codes phonétiques par bloc de 60 caractères maximum,
- réponse quasi immédiate après la réception du dernier caractère,
- décodage automatique des nombres entiers positifs ou négatifs inférieurs à 1 milliard,
- possibilité de répétition programmée de l'édition de 2 à 9 fois,
- commutateur pour fonctionnement en test (autonome),
- commutateur de sortie: vocale, imprimée, ou vocale plus imprimée.

REFERENCES

(1) - E. LEIPP, J.S. LIENARD, M. CASTELLENGO, J. SAPALY, D. TEIL, A. CALINET, M. MLOUKA - Colloque sur la parole. Bulletin du Groupe d'Acoustique musicale de l'Université de Paris VI n° 53, janvier 1971.

(2) - J. QUINIO, D. TEIL - La synthèse de la parole par ordinateur à partir de digrammes phonétiques. Revue d'Acoustique t.3, n° 9, 1970.

(3) - E. LEIPP, M. CASTELLENGO, J.S. LIENARD, J. QUINIO, J. SAPALY, D. TEIL - Générateur synthétique de parole. Brevet ANVAR n° 1602936, février 1971.

(4) - J.S. LIENARD, D. TEIL - Les éléments phonétiques et la traduction automatique du message écrit en message parlé - Automatisme n° 10, Octobre 1970.

COMMANDE D'UN SYNTHETISEUR A FORMANTS PAR ORDINATEUR

M. MRAYATI - E.N.S. d'Electronique & de Radioélectricité de GRENOBLE

RESUME :

Un synthétiseur à formants de type série est commandé par un ordinateur. Le système de contrôle utilise les périphériques d'entrée-sortie graphique. Un filtre numérique câblé effectue le lissage des paramètres. Deux programmes permettent les communications entre l'opérateur, l'ordinateur et le synthétiseur ainsi qu'avec les autres périphériques de stockage et d'entrées/sorties. L'ordinateur est un PDP.11 avec 12 K de mémoire centrale. La combinaison ordinateur-synthétiseur permet une synthèse de la parole en temps réel sous le contrôle d'un programme. L'opérateur peut successivement synthétiser une phrase, la juger, la modifier et l'écouter à nouveau rapidement.

SUMMARY :

A formants series synthesizer is digitally controlled from a computer. This control system makes use of graphics input/output devices. Hardware digital filtering of the control parameters is incorporated.
Two programs handle all communications between the operator, the computer and the synthesizer, as well as other storage and input/output devices. The computer is a PDP.11 with 12 K memory. The computer-synthesizer combination permits on-line, real-time synthesis of speech under program control. The experimenter exercises control over the synthesizer with a series of commands typed on a display console. By this method, an utterance may be synthesized, judged, modified and tried again in rapid succession.

COMMANDE D'UN SYNTHETISEUR A FORMANTS PAR ORDINATEUR

M. MRAYATI
E.N.S.E.R. GRENOBLE

1. INTRODUCTION

Le synthétiseur à formants, dans son principe, simule la fonction de transfert du conduit vocal. Ce type de synthétiseur permet d'obtenir une parole de bonne qualité. Sa réalisation pratique est devenue très aisée avec les derniers développements des composants électroniques [1].

La commande par un mini-ordinateur permet d'utiliser au mieux toutes les possibilités d'un tel système de synthèse. La commande par un ensemble entièrement électronique tel que le lecteur de courbe par caméra de télévision n'a que des possibilités limitées et il est difficilement modifiable. Par contre, si tout le système est simulé sur ordinateur, la commande du synthétiseur sera plus souple, mais la synthèse ne pourra pas être effectuée en temps réel. En conséquence, il serait nécessaire de disposer, soit d'une mémoire centrale très importante, soit d'une mémoire de masse d'accès rapide, permettant d'emmagasiner l'onde de sortie. De plus, nous pensons qu'une méthode efficace de synthèse paramétrique est celle où l'expérimentateur peut entendre immédiatement l'effet de ses commandes. Ce résultat peut être obtenu en utilisant harmonieusement la conjugaison des possibilités "hardware" d'un synthétiseur câblé et "software" d'un ordinateur. Dans ce cas, un mini-ordinateur (avec 4 K.mots de mémoire centrale) est suffisant pour contrôler de façon souple et précise le synthétiseur à formants.

L'emploi de périphériques tels que des organes d'entrée-sortie graphiques facilite les liaisons entre l'expérimentateur, ses données et le système de commande.

Notre système de commande comporte deux parties :

. une partie HARDWARE composée de deux organes : d'une part une unité d'acquisition et de visualisation graphique des courbes images de l'évolution avec le temps des paramètres de commande, d'autre part un ensemble réalisant la liaison ordinateur-synthétiseur et comportant coupleur, filtre numérique et convertisseur D/A ;

. une partie SOFTWARE comprenant deux programmes : un programme permet de réaliser l'acquisition des paramètres à partir des tracés, sonagrammes par exemple, et crée une bibliothèque ; un autre programme gère les communications entre l'opérateur, l'ordinateur et le synthétiseur.

2. PRESENTATION GENERALE DU SYSTEME

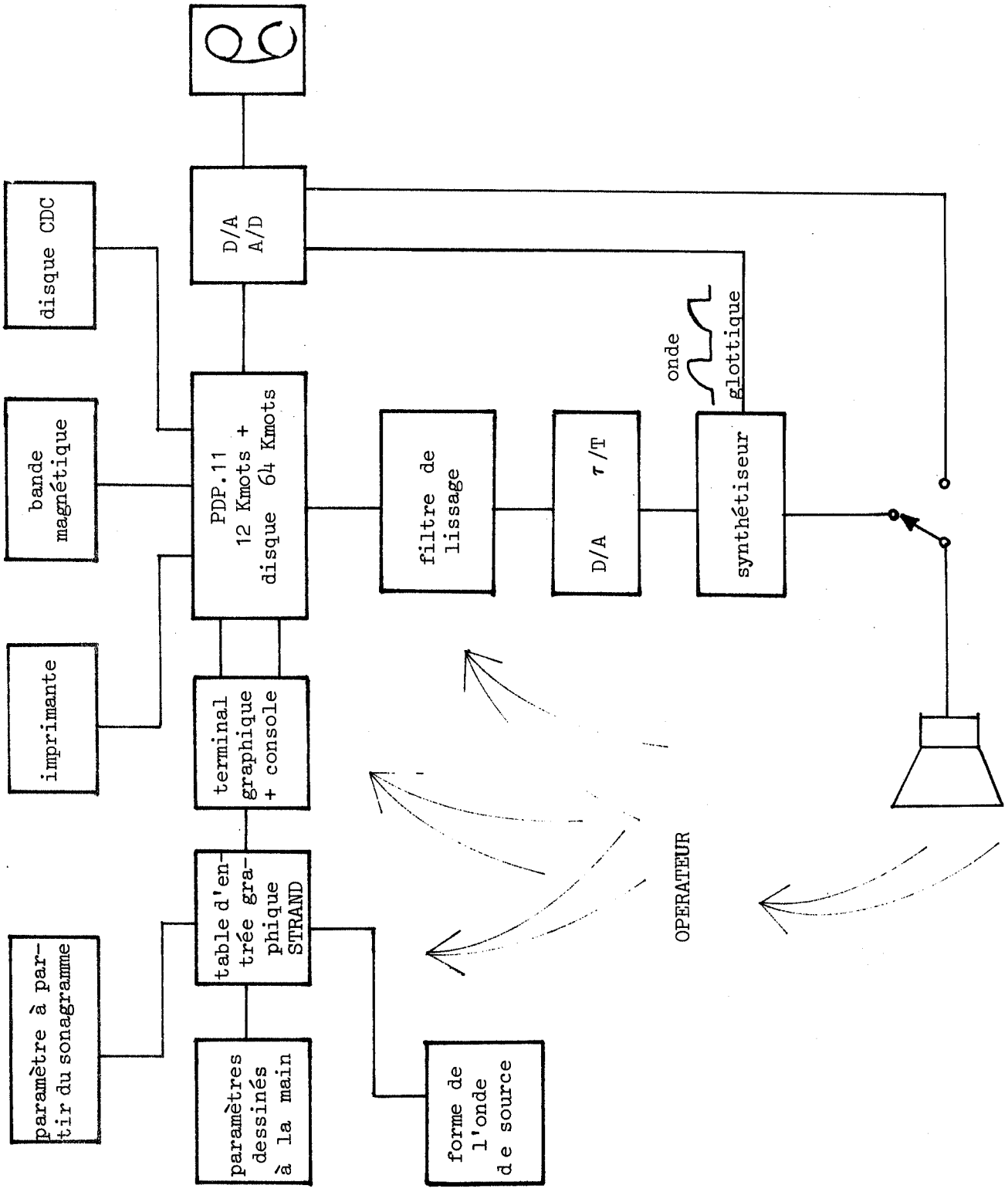
La figure 1 (page 3) montre le schéma synoptique du système qui comprend :

Un ordinateur PDP.11/20 ayant une mémoire centrale de 12 K.mots de 16 bits et un disque de 64 K.mots. Le stockage des fichiers de synthèse peut être effectué, soit par le disque C.D.C. (2 500 000 mots), soit par unité de bandes magnétiques.

Le contenu des fichiers de synthèse peut être imprimé. Chaque ligne de cette table donne les valeurs des paramètres, exprimées en Hz et en dB, à un instant déterminé, la période d'échantillonnage étant de 20 ms.

Une console de visualisation, dont l'écran est un tube à mémoire, est utilisée comme clavier et comme terminal d'entrée/sortie graphique. A ce terminal graphique-alphanumérique, est couplée une table STRAND, Système de TRANscription de Données, qui comprend une plaque de verre recouverte d'une couche conductrice transparente et un crayon relié à un ensemble électronique. Cette table délivre les coordonnées du point désigné à l'aide du crayon par l'opérateur. L'échelle de correspondance entre l'écran du terminal graphique et la table STRAND peut être modifiée. Un programme d'entrée/sortie graphique utilisant la configuration décrite ici, a été mis au point [2]. Cet ensemble permet l'acquisition et la visualisation des données graphiques.

FIGURE 1.


disque CDC
D/A
A/D
bande magnétique
PDP.11
12 Kmots +
disque 64 Kmots
filtre de lissage
D/A τ/T
onde glottique
synthétiseur
imprimante
terminal graphique + console
table d'entrée graphique STRAND
paramètre à partir du sonagramme
paramètres dessinés à la main
forme de l'onde de source
OPERATEUR

Le synthétiseur à formants est relié à l'ordinateur au moyen d'une interface composée d'un coupleur, d'un filtre numérique et d'un convertisseur D/A. Le coupleur fait la liaison entre l'ordinateur et l'entrée du filtre numérique, fournit les valeurs des paramètres toutes les 20 ms et transmet "une interruption" à l'ordinateur. Le filtre numérique effectue le lissage des paramètres. Au niveau du synthétiseur, pour certaines expériences, on peut commander manuellement un ou plusieurs paramètres, les autres continuant à être contrôlés par l'ordinateur.

Nous disposons également d'une unité d'acquisition de donnée comportant des convertisseurs A/D et D/A. Elle peut être utilisée pour fournir au synthétiseur un signal de source vocale en même temps que les autres paramètres. La forme de l'onde glottique peut être acquise à partir de la table STRAND.

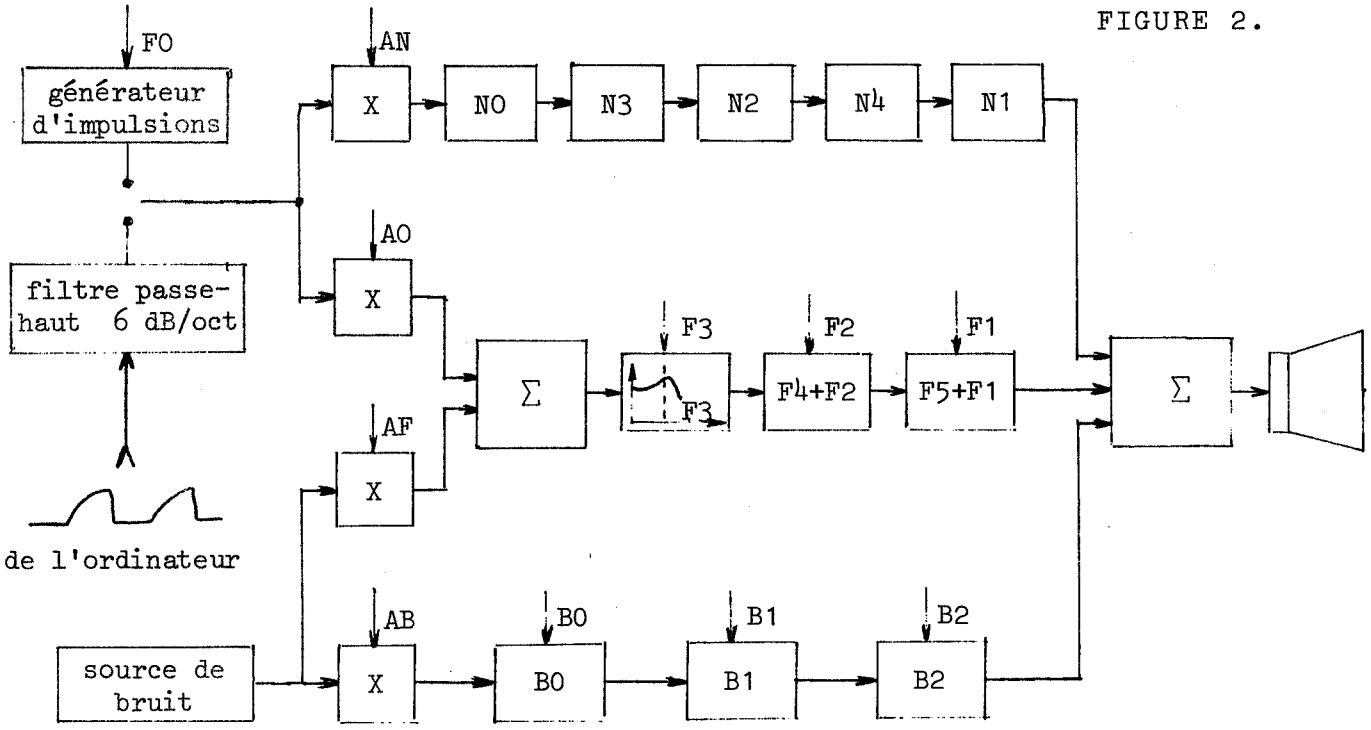
3. DESCRIPTION DU SYNTHETISEUR ET DU FILTRE NUMERIQUE

3.1. LE SYNTHETISEUR COMMANDÉ [1]

La structure du synthétiseur est donnée figure 2 (page 5). Le générateur d'impulsions, image de la source vocale, dont la fréquence fondamentale est commandée par FØ, attaque deux canaux différents. Le premier simule le conduit nasal et il comprend des circuits introduisant des zéros et des pôles, l'amplitude étant commandée par AN. Le second canal représente le conduit vocal ; l'amplitude est ici commandée par AC et les divers circuits de formants sont commandés par F1, F2, F3. Le troisième canal, canal de bruit, est constitué par des circuits de formants et d'antiformants commandés par B1, B2, B0, l'amplitude l'étant par AB. L'amplitude de bruit introduit dans le canal vocal est commandée par AF.

On voit que les paramètres à commander sont au nombre de 11, mais pour effectuer des recherches fondamentales sur la parole, il peut être intéressant de commander également, par exemple, les largeurs des bandes passantes des circuits de formants ; aussi avons-nous prévu, au plus, 16 paramètres de commande.

FIGURE 2.



3.2. LES PARAMETRES DE COMMANDE

La fréquence d'échantillonnage des paramètres est de 50 Hz. Les paramètres sont codés avec 8 bits. La quantité d'information nécessaire pour synthétiser des phrases avec 10 paramètres est de 4 000 bits/s. On en déduit que 250 mots de 16 bits sont suffisants pour coder une seconde de parole quelconque.

Les spécifications adoptées pour le codage des paramètres mémorisés sont données au tableau 1, page 6. Les correspondances d'échelle pour l'acquisition à partir de la table d'entrée graphique sont adaptées aux échelles d'un sonagramme. De ces paramètres, on peut obtenir une résolution plus fine avec des tracés ayant des échelles convenables.

TABLEAU 1. PARAMETRES DE CONTROLE ET LEURS SPECIFICATIONS

paramètre n°	paramètre	plage de variation adoptée	échelle sur table STRAND	quantification	incrément	remarques
1	FØ	50 - 305 Hz	50 Hz/cm	8 bits	1 Hz	fréquence fondamentale
2	F1	100 - 1120 Hz	500 "	7 "	4 Hz	premier formant
3	F2	500 - 3050 Hz	500 "	8 "	10 Hz	deuxième formant
4	F3	1000 - 4570 Hz	500 "	8 "	14 Hz	troisième formant
5	B1	1000 - 6100 Hz	1000 "	8 "	20 Hz	premier formant de bruit
6	B2	2000 - 10 670 Hz	1000 "	8 "	34 Hz	deuxième formant de bruit
7	AØ	0 - 32 dB	8 dB/cm	5 "	0,25 dB	amplitude de la source vocale
8	AN	0 - 32 dB	8 "	5 "	0,25 dB	amplitude de nasalité
9	AB	0 - 32 dB	8 "	5 "	0,25 dB	amplitude du bruit
10	AF	0 - 32 dB	8 "	5 "	0,25 dB	amplitude du bruit dans les formants vocaux
. . . . 16						addition optionnelle de six paramètres

3.3. FILTRE NUMERIQUE DE LISSAGE DE PARAMETRE [3]

Le développement rapide des circuits intégrés rend les filtres numériques plus intéressants, pour certaines applications, que les filtres analogiques aux points de vue précision, stabilité, facilité de commande, prix et dimensions. Un autre avantage du filtre numérique est d'offrir une possibilité de multiplexage. Avec le même circuit (unité arithmétique), on peut :

a. filtrer plusieurs signaux en même temps ;

b. filtrer un même signal avec des caractéristiques différentes ;

c. faire une combinaison des deux fonctions précédentes.

Nous avons adopté la dernière configuration pour le filtrage de nos 16 paramètres, multiplexés, avec possibilité de changer en temps réel les fréquences de coupure de chaque paramètre.

Nous avons réalité le filtre passe-bas dont la fonction de transfert dans le plan Z est :

$$H(z) = \frac{(1 - e^{-T/\tau})^2 \; z^{-1}}{(1 - e^{-T/\tau} \; z^{-1})^2}$$ qui peut être écrit sous la forme $$H(z) = \frac{L_1 z^{-1}}{1 - K_1 z^{-1} - K_2 z^{-2}}$$.

L'équation aux différences est donc :

$$Y(nT) = K_2 Y(nT-2T) + K_1 Y(nT-T) + L_1 X(nT-T)$$

où X(nT) est le signal d'entrée et Y(nT) le signal filtré.

Le circuit utilisant le plus faible nombre possible d'éléments électroniques est donné figure 3.

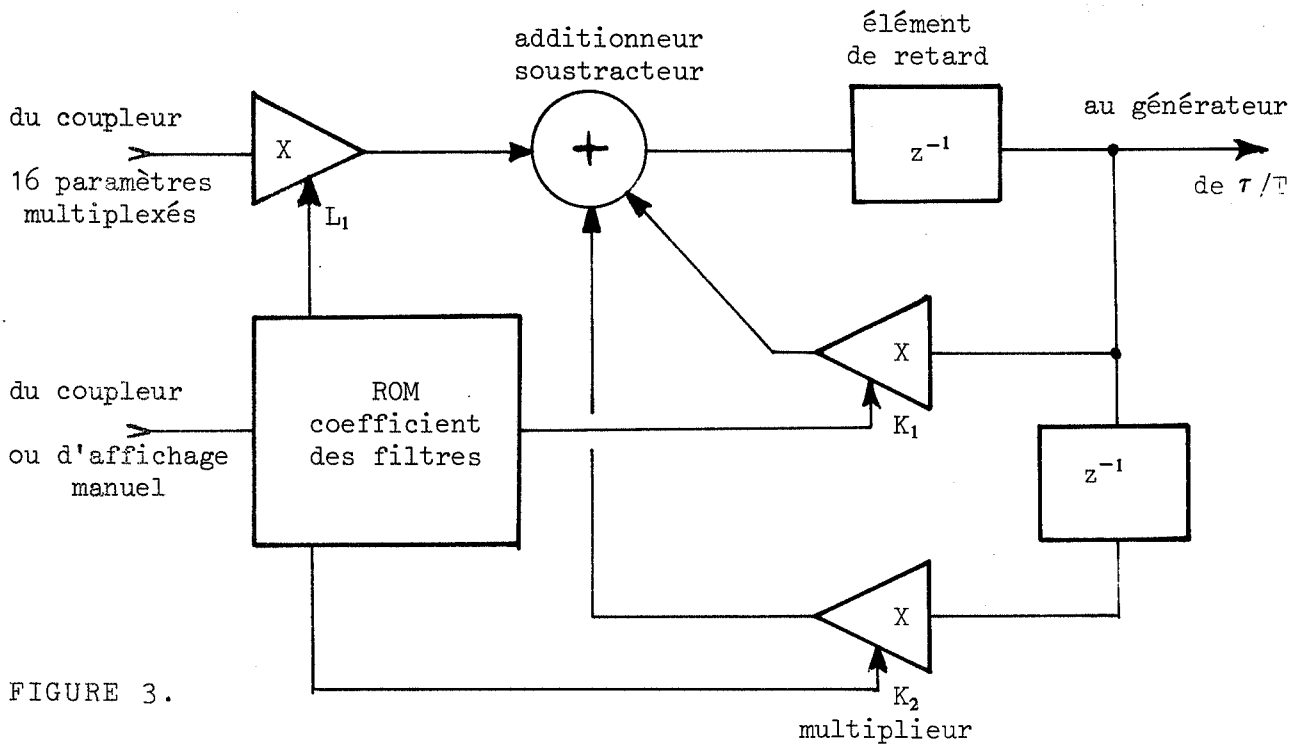


FIGURE 3.

Le calcul dans l'unité arithmétique du filtre est effectué sous forme série avec 16 bits. L'entrée et la sortie du filtre sont données avec 8 bits. La fréquence d'échantillonnage du filtre est de 1000 Hz. La commande des coefficients des filtres est réalisée par un adressage de 3 bits d'une mémoire morte donnant 8 possibilités de fréquence de coupure, les coefficients étant fournis avec 16 bits.

4. DESCRIPTION DE L'ENSEMBLE DES PROGRAMMES DE GESTION

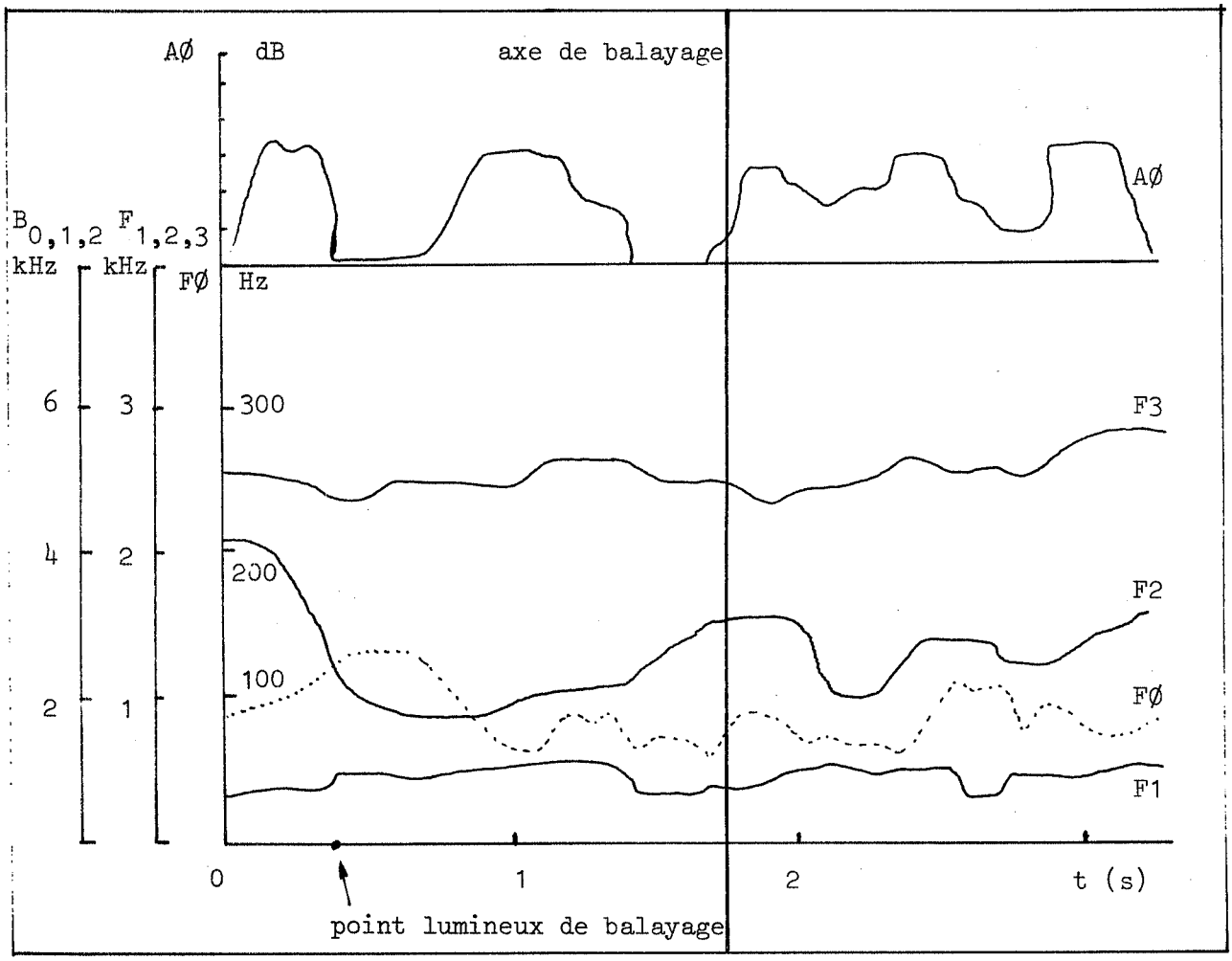
4.1. PROGRAMME D'ACQUISITION

Ce programme sert à construire la bibliothèque des synthèses. Il réalise les acquisitions, à partir de la table STRAND, des courbes images de l'évolution dans le temps des paramètres de la parole à synthétiser. Ces courbes, échantillonnées toutes les 20 ms, sont reproduites, au fur et à mesure de l'acquisition, sur l'écran du terminal graphique. Le programme est conversationnel. A partir de la table STRAND et du clavier, l'utilisateur peut effectuer les opérations suivantes :

. spécifier le temps de synthèse en ms ;
. spécifier le nom sous lequel la phrase sera stockée ;
. effacer l'écran et reproduire des axes gradués automatiquement ;
. achever l'acquisition d'une courbe en la traçant à l'aide du crayon et de la table STRAND et visualiser l'opération en même temps sur l'écran ;
. modifier une certaine courbe ;
. imprimer la table des paramètres acquis en Hz et en dB pour avoir une copie numérique des courbes échantillonnées ;
. coder tous les paramètres en 8 bits, forme sous laquelle ils sont stockés (chaque échantillon de 10 paramètres dans 5 mots de 16 bits) ;
. imprimer la table des paramètres après le codage ;
. répéter ou terminer le déroulement du programme.

Le programme est écrit en langage assembleur et il occupe approximativement 2 000 mots de 16 bits. La figure 4 (page 9) montre un exemple de visualisation des paramètres sur l'écran du terminal graphique.

FIGURE 4.



4.2. PROGRAMME DE CONTRÔLE

C'est le programme qui gère toutes les communications entre l'opérateur, l'ordinateur et le synthétiseur ainsi que la bibliothèque des phrases déjà stockées et le système d'entrée/sortie graphique. Les commandes frappées au clavier par l'opérateur sont composées de deux caractères. Le programme comprend plusieurs subroutines chargées d'effectuer les différentes fonctions désirées. Il commence par réserver un buffer "espace de travail" et puis, il efface l'écran du terminal graphique et affiche des axes gradués. Il cherche dans la bibliothèque une phrase test qui permette de voir si le système fonctionne, puis il rend le contrôle à l'opérateur qui peut effectuer les opérations suivantes :

★ appeler une phrase déterminée de la bibliothèque dans le buffer du programme "espace de travail". Cette opération est faite en affichant à partir du clavier le nom du fichier correspondant. Le programme effectue ce travail puis efface l'écran et trace les axes sur l'écran ;

★ visualiser l'évolution avec le temps d'un ou plusieurs paramètres. Par exemple, l'utilisateur peut demander le dessin de F0, F1, F2, F3 et A0 ou de BA, B2 et AB ... etc. ;

★ commander le synthétiseur en lui fournissant les paramètres préalablement appelés dans une zone de mémoire (espace de travail). Automatiquement, les valeurs des paramètres sont envoyées au synthétiseur toutes les 20 ms et un point lumineux balaie l'axe des temps sur l'écran, indiquant l'abscisse des échantillons alors fournis par l'ordinateur ;

★ commander le synthétiseur avec un balayage manuel de l'axe des temps. Le programme affiche sur l'écran un axe vertical dont la position peut être contrôlée par l'utilisateur à partir du crayon de la stable STRAND ;

★ modifier des points ou des parties de l'évolution d'un paramètre. Le programme affiche des axes perpendiculaires dont le point d'intersection peut être positionné à partir du crayon de la table STRAND. Ces coordonnées peuvent remplacer les anciennes valeurs de la courbe choisie ;

★ re-stocker la phrase dans la bibliothèque sous un autre nom ou sous le même nom après avoir effacé l'original ;

★ fournir une certaine forme d'onde glottique acquise à partir de la table STRAND ;

★ terminer le programme.

Chaque opération peut être répétée autant de fois qu'il est désirable. Ce programme est écrit en langage assembleur et sa taille est d'environ 2 300 mots de 16 bits.

5. APPLICATION

Le système de contrôle du synthétiseur à formants décrit ici permet d'effectuer de nombreuses études sur la production de la parole.

Des messages parlés peuvent être synthétisés à partir de mots isolés, préalablement analysés et dont les paramètres ont été stockés.

Un système de synthèse par dyades (ou diphonèmes) est à l'étude. Le but est de synthétiser un message parlé à partir d'un message écrit.

Ce système est bien adapté pour le stockage de la parole sous forme paramétrique. La capacité occupée par une seconde de parole codée à l'aide de 10 paramètres et ayant une bonne qualité est, comme nous l'avons vu, de 250 mots de 16 bits.

L'étude du rôle des paramètres, des transitions, de l'effet de la forme de l'onde de source vocale au niveau de la perception est rendue très aisée.

6. CONCLUSION

Nous n'avons pas cherché dans ce système à avoir le minimum de quantité d'information de commande. Le but est d'avoir un outil pratique pour effectuer nos recherches.

Dans la conception de cette commande, nous avons visé deux aspects :

1. aucune connaissance de la programmation ou de la manipulation de l'ordinateur n'est nécessaire pour l'utiliser,
2. l'expérimentateur n'a pas de messages complexes à échanger avec l'ordinateur ; il suffit de connaître un code de commande très simple.

Les programmes sont composés d'un ensemble de subroutines très modulaires, ce qui permet d'effectuer facilement des modifications.

BIBLIOGRAPHIE

[1] *J. PAILLE*
Contribution aux études sur la synthèse paramétrique de la parole, synthétiseur à formants, analogue de la source vocale.
Thèse de Docteur ès-Sciences physiques. Grenoble (1971)

[2] *M. MRAYATI*
Manuel d'utilisation d'entrée/sortie graphique
Rapport interne. E.N.S.E.R. Grenoble (juillet 1973)

[3] *M. MRAYATI*
Interface synthétiseur de parole, ordinateur comprenant filtre numérique
Rapport D.E.A. E.N.S.E.R. Grenoble (juillet 1971)

Thème n° 3

APPLICATION DES CONTRAINTES LINGUISTIQUES A LA RECONNAISSANCE AUTOMATIQUE DE LA PAROLE

Recherche lexicale par utilisation de contraintes phonétiques en reconnaissance analytique de la parole

J. -P. HATON

Laboratoire d'Electricité et d'Automatique

Université de Nancy 1

RESUME

L'approche analytique de la reconnaissance de la parole donne d'un mot prononcé une chaîne phonémique plus ou moins entachée d'erreurs. On s'intéresse ici à la recherche lexicale permettant de passer de cette chaîne au mot qui a été prononcé. On propose un algorithme de comparaison dynamique qui fournit une solution efficace à ce problème, même dans le cas de réponses multiples. Cet algorithme utilise largement les contraintes phonétiques par l'utilisation des variantes possibles de la transcription phonétique d'un mot et par le recours à une "matrice des proximités" de phonèmes déterminée par apprentissage. Le système fonctionne en temps réel sur mini-calculateur T 2000.

SUMMARY

Analytical speech recognition consists of translating any utterance into an error-full string of phonemes. In this paper we describe a lexical search procedure which permits to match such strings with reference transcriptions of words. The system uses a dynamic matching algorithm and operates as well with multiple-labelled segments strings. It includes phonetic constraints by the use of graph representation of words and of a "proximity matrix" for the computation of interphonemes distances. Since an exhaustive search is quite unpractical for large vocabularies, several methods are used to restrict the comparison to those words which are close -in some sense- to the incoming string.

I - Introduction.

La réponse d'un système de reconnaissance analytique de la parole peut être mise sous forme d'une chaîne phonémique plus ou moins entachée d'erreurs, qui correspond à la transcription, effectuée par le système, du mot ou de la phrase émise par un locuteur. Pour certaines applications -transmission de la parole par exemple- cette chaîne phonémique suffit ; mais le plus souvent il est nécessaire de retrouver la forme du message qui a été prononcé.

Des analyses lexicale, morphologique ou syntaxique permettent, suivant les cas, d'effectuer ce travail, rendu délicat par les erreurs de répétition, élision, insertion et/ou confusion de phonèmes introduites au niveau de la segmentation et de la reconnaissance. On s'intéresse ici au seul problème de la recherche lexicale consistant à retrouver un mot dans un vocabulaire donné, à partir de la chaîne phonémique. La méthode proposée est une comparaison dynamique de la chaîne phonémique étudiée et des transcriptions phonétiques des mots du lexique. Nous avons déjà utilisé une méthode analogue en reconnaissance acoustique globale de mots isolés, elle permet en particulier ici, de comparer efficacement des chaînes de phonèmes dont les longueurs diffèrent, même de façon notable.

Cette méthode est appliquée à la reconnaissance lexicale de la parole, elle serait également valable dans tout problème de reconnaissance dans lequel les formes à comparer peuvent être décrites sous forme d'une suite ordonnée d'éléments.

II - Principe de l'algorithme de comparaison dynamique.

Si l'on admet la possibilité de réponses multiples de la part de l'étage de reconnaissance acoustique, un mot prononcé sera décrit par une chaîne de phonèmes C :

$$C = \left\{ (c_1^1 , c_2^1) , (c_1^2 , c_2^2) , \ldots , (c_1^n , c_2^n) \right\}$$

dans le cas de réponses doubles ; c_1^j et c_2^j représentent les phonèmes les plus ressemblants au i^e phonème du mot prononcé, classés respectivement en première et deuxième positions. Chacun des éléments de la chaîne C possède par ailleurs un "score acoustique" A_k^j , k = 1, 2, qui précise la réponse du niveau acoustique.

Il s'agit d'identifier la chaîne C par comparaison séquentielle aux N mots du lexique représentés par leur transcription phonétique :

$$R_p = \left\{ r^1, r^2, \ldots, r^m \right\} \quad , p = 1, 2, \ldots, N$$

En réalité, ces chaînes R_p rendent compte également des contraintes phonétiques concernant les mots du lexique (différentes variantes de prononciation) : chacune est constituée d'un arbre dans lequel apparaissent les diverses variantes d'élision, d'insertion ou de substitution de phonèmes constatées pendant un apprentissage. Cette représentation compacte est beaucoup plus intéressante que celle qui consisterait à avoir autant de transcriptions pour un mot que de variantes possibles de ce mot.

Cette comparaison résulte en un taux de similitude $S_p = S(R_p, C)$ tel que

$$S(R_p, C) = 0 \iff C \equiv R_p$$

Plusieurs problèmes se posent dans le calcul de S_p, concernant en particulier les différences de longueur des chaînes, et le nombre apparemment élevé de comparaisons à effectuer. L'algorithme de comparaison dynamique suivant fournit une solution efficace :

étant donnée une métrique d interphonèmes, la condition initiale

$$s(1) = \underset{k=1,2}{\text{Min}} \left\{ d(r^1, c_k^1) \right\}$$

et la relation de récurrence

$$s(\ell+1) = s(\ell) + \underset{k=1,2}{\text{Min}} \begin{cases} d(r^{i+1}, c_k^j) \\ d(r^{i+1}, c_k^{j+1}) \end{cases}$$

-si l'on admet que la $\ell^{ème}$ comparaison avait abouti aux éléments r_i et c^j_k des deux chaînes- conduisent finalement à un taux de similitude global

$$S(R_p, C) = \frac{1}{L} s(L)$$

où L est le nombre de comparaisons effectuées.

La distance interphonèmes d est calculée à partir d'une "matrice des proximités" de phonèmes, déterminée par apprentissage sur le système de reconnaissance acoustique de façon analogue à une matrice de confusion. Cette façon de procéder permet de tenir compte au mieux des contraintes phonétiques car d dépend de la proximité phonétique des phonèmes et aussi des performances du reconnaisseur acoustique.

Cet algorithme définit un chemin optimal de comparaison des chaînes dans l'espace à trois dimensions construit sur R_p, C_1 et C_2. Un exemple pratique de chemin de comparaison est donné fig. 1.

III - Résultats et discussion.

Cette méthode de recherche lexicale a été testée pour l'instant sur un vocabulaire d'une quarantaine de mots, avec un système de reconnaissance analytique déjà décrit par ailleurs [1]. Les résultats sont fortement tributaires des performances de l'étage de reconnaissance acoustique. Voici quelques exemples de résultats :

$\{(/s/,/\int/),(/s/,/z/),(/a/,/ə/),(/o/,/a/),(/a/,/o/),(/b/,/p/),(/a/,/\hat{\varepsilon}/),(/\hat{\varepsilon}/,/ã/)\}$

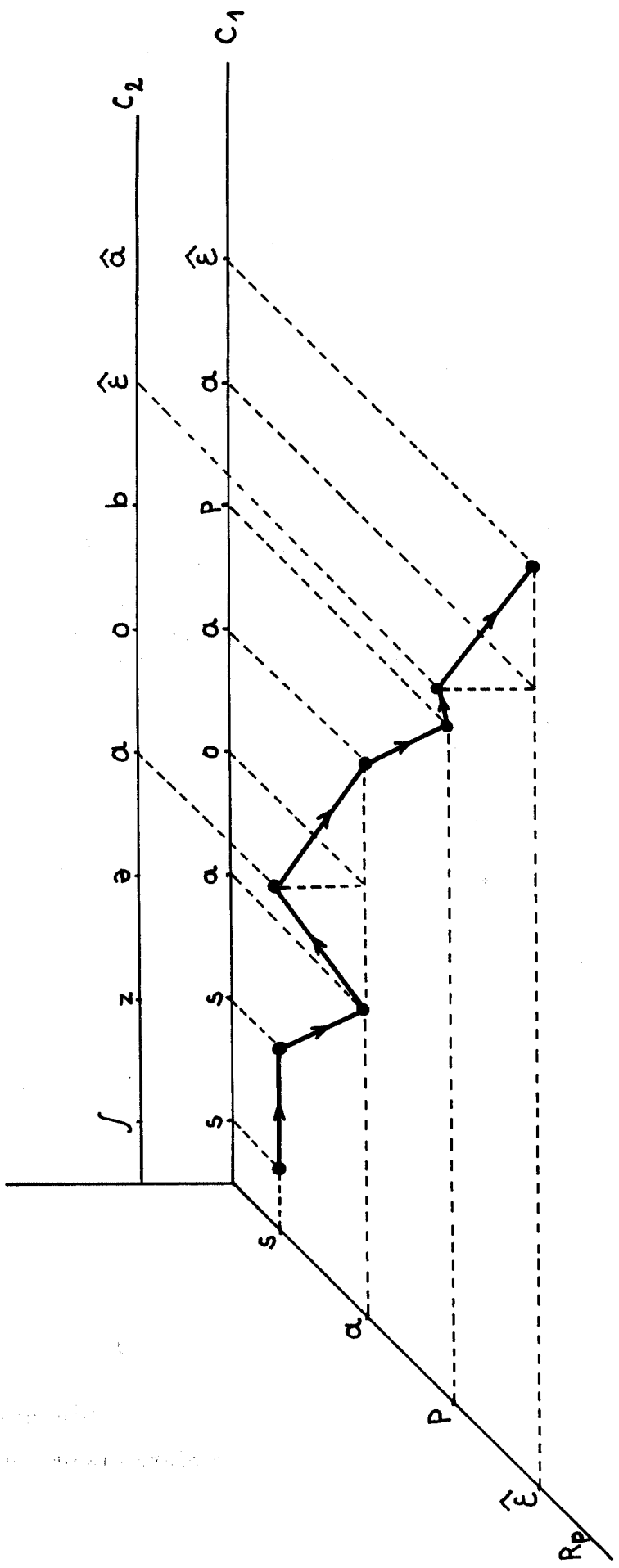
$\longrightarrow$ SAPIN

$\{(/i/,/u/),(/o/,/i/),(/b/,/p/),(/o/,/\ell/),(/u/,/o/),(/s/,/ʒ/),(/ʒ/,/z/),(/\int/,/s/)\}$

$\longrightarrow$ EPOUSE

$\{(/a/,/ø/),(/e/,/i/),(/p/,/t/),(/r/,/\ell/),(/ə/,/a/),(/i/,/e/)\}$ $\longrightarrow$ APPÂT

L'algorithme de comparaison dynamique fournit un chemin optimal dont la fig. 1 fournit un exemple, dans le cas de la première chaîne qui vient d'être citée et de $R = \{/s/,/a/,/p/,/\hat{\varepsilon}/\}$.



Exemple pratique de chemin de comparaison

Fig 1

Le temps de comparaison d'une chaîne inconnue à un mot de référence varie en moyenne de 1 à 2 ms, selon la longueur, sur un minicalculateur Télémécanique T 2000, relativement peu rapide. Ces performances permettent d'envisager l'utilisation de vocabulaires importants (plusieurs centaines de mots), avec réponse quasi-immédiate.

Dans ce dernier cas, il est de toute façon intéressant d'optimiser le temps de calcul en restreignant la comparaison d'une chaîne aux seuls mots "ressemblant" -en un certain sens- à cette chaîne. Pour cela, il faut mettre au point des méthodes rapides de mesure de "voisinage" entre chaînes. Jusqu'à présent, les critères que nous avons utilisés concernent :

- l'examen de la longueur des chaînes : ce critère est intuitif, dans la mesure où un mot de quatre phonèmes, par exemple, a peu de chances d'être transcrit en une chaîne de dix éléments si l'étage de reconnaissance acoustique fonctionne de façon acceptable. Avec les notations du paragraphe II, le critère

$$\frac{m}{2} \leqslant n \leqslant 2m$$

s'est révélé satisfaisant dans tous les cas.

- la partition du vocabulaire en sous-vocabulaires constitués de mots assez semblables phonétiquement (contenant des occlusions, des bruits de friction, etc..., ces propriétés étant assez faciles à détecter sur les mots prononcés).

On peut envisager des méthodes à la fois plus rapides et plus souples, par exemple par recherche associative sur certains "phonèmes-clés" tels que fricatives ou occlusives.

IV - Bibliographie.

[1] J.-P. HATON "Contribution à l'analyse, la paramétrisation et la reconnaissance automatique de la parole" Thèse d'Etat, Université de Nancy I, 8 janvier 1974.

PREDICTION DE MOTS PAR CONTRAINTES PHONOLOGIQUES

L. MICLET

ECOLE NATIONALE SUPERIEURE DES TELECOMMUNICATIONS
46 RUE BARRAULT - PARIS 13 ème

En Reconnaissance Automatique de la parole, après les étapes d'analyse, de paramétrisation, de segmentation et de reconnaissance phonémique, le mot à reconnaître se présente sous la forme d'un tableau de "candidats-phonèmes" : à chacun des segments est affectée une liste de phonèmes prédits, rangée par ordre de confiance décroissant. Une combinaison des candidats-phonèmes fournit un "candidat-mot", dont le dictionnaire du système décidera l'appartenance ou non au vocabulaire. Un assez grand nombre de candidats-phonèmes pour chaque segment assure que le phonème à reconnaître fait partie des candidats proposés, mais le nombre de recherches dans le dictionnaire devient alors trop important. On présente ici une méthode de prédiction de mots, c'est-à-dire de réduction a priori du nombre de candidats-mots, en contraignant les candidats-phonèmes à s'enchaîner selon des règles phonologiques imposées. Le gain de recherche dans le dictionnaire semble être assez important pour justifier l'introduction de telles méthodes de prédiction, du moins dans des systèmes de reconnaissance travaillant sur de gros dictionnaires.

WORD PREDICTION USING PHONOLOGICAL CONSTRAINTS

L. MICLET

Once one has gone through the phoneme analysis, parametrisation, segmentation and recognition phases in automatic speech recognition, the word to be recognized has been transformed into a tableau ; this tableau is obtained by listing for each phoneme position in the word all predicted phoneme candidates for recognition in that position, as ranked by decreasing confidence values. Any combination made of one phoneme-candidate in each position yields a word-candidate ; by comparing it to the content of the system's dictionnary, it will be decided whether this word-candidate belongs to the vocabulary or not. By selecting a fairly large number of phoneme-candidates for each position, we may guarantee that the actual phoneme (hopefully to be recognized) belongs to the above mentioned related list ; however, the number of search procedures to achieve within the dictionnary may then become too large.

In this paper, we present a word prediction method, i. e. an approach whereby the number of word-candidates to seek after is reduced a priori. This reduction can be obtained by constraining the phoneme-candidates to be chained together according to specified phonological rules. The gain for the seeking process within the dictionnary seems to be sufficient to justify the implementation of such prediction methods into speech recognition systems working on large dictionnaries.

RECONNAISSANCE DE LA PAROLE :

PREDICTION DE MOTS PAR CONTRAINTES PHONOLOGIQUES

L. MICLET

LABORATOIRE D'AUTOMATIQUE, NUMERIQUE
ET ANALOGIQUE de l'Ecole Nationale
Supérieure des Télécommunications

Introduction

La méthode présentée ici est destinée à prendre place dans un système de reconnaissance analytique de la parole, travaillant sur de gros dictionnaires. Elle présente une méthode de prédiction des mots, sur des critères de nature phonologique, afin de limiter le nombre de recherches lexicales. Le modèle de prédiction est un automate fini dont le vocabulaire terminal est constitué en première approximation des phonèmes de la langue française. L'analyse d'une suite de phonèmes par cet automate permet de prédire si l'on doit ou non la rechercher dans le dictionnaire : l'automate rejette a priori un grand nombre de chaînes de phonèmes, dont il sait qu'elles ne représentent pas des mots du vocabulaire.

Reconnaissance phonémique

Un modèle analytique de reconnaissance de la parole suppose que l'on se soit fixé une unité de segmentation et de reconnaissance plus petite que le mot. L'unité choisie ici est le phonème. Une émission de voix à reconnaître sera donc, après paramétrisations et normalisations, segmentée en unités les plus voisines possibles du phonème.
La reconnaissance consiste alors à calculer pour chaque segment une "distance" aux phonèmes de la langue, dont les paramètres ont été déterminés par apprentissage. A la suite de cette étape, un mot peut alors se représenter sous la forme d'un tableau, dont chaque ligne représente la reconnaissance d'un segment : elle est constituée d'une liste de phonèmes

de la langue, chacun étant flanqué d'un taux de confiance, calculé à partir de la "distance" de ce phonème au segment à reconnaître. Au sens de la reconnaissance phonémique du système, le phonème ayant le plus fort taux de confiance dans une ligne donnée est le "candidat" le plus probable pour le segment correspondant. Le mot à reconnaître peut alors se décrire comme un chemin descendant dans le tableau, une suite de "candidats-phonèmes" affectés de taux de confiance. Du point de vue de la prédiction, tout chemin de ce type est un "candidat-mot". Le problème est donc de sélectionner le bon candidat-mot, ou plusieurs candidats-mots, avant de s'élever dans les niveaux grammaticaux ou éventuellement sémantiques du système de reconnaissance. Il est certain qu'un assez grand nombre de candidats-phonèmes par segment assure la présence du bon candidat parmi les combinaisons des candidats-phonèmes. Une première prédiction consistera donc à optimiser le nombre de candidats-phonèmes par segment, de façon à d'une part limiter le nombre de candidats-mots issus du tableau, d'autre part à éviter de supprimer le bon candidat-phonème dans la ligne affectée au segment correspondant.
La détermination de paramètres permettant de calculer ce nombre optimal fait partie de l'apprentissage du système. Elle est essentiellement liée à la façon de séparer les phonèmes, et de calculer le taux de confiance. On ne saurait donc proposer de règles générales pour la prédiction au niveau des phonèmes. A titre d'exemple, pour notre système, en moyenne 4 candidats sont retenus par ligne, et la probabilité d'exclusion du bon candidat est inférieure à 5 %.

Prédiction des candidats-mots

1) Choix du modèle

Sur l'ensemble des candidats-mots que l'on peut former à partir du tableau issu de la première prédiction, on peut maintenant tenter de prédire ceux qui n'appartiennent certainement pas au dictionnaire, avant toute recherche lexicale, afin de diminuer le nombre de fois où ce dictionnaire sera parcouru en vue de vérification : le dictionnaire décidera de l'appartenance ou non au vocabulaire du candidat-mot accepté par la prédiction.

Nous avons choisi de retenir des critères d'élimination de candidats-mots qui sont de nature phonologique, de façon à utiliser le plus tôt possible dans la hiérarchie du système de reconnaissance, le caractère linguistique de l'unité choisie : le phonème. D'autre part, ce type de prédiction est naturellement non lié au système de reconnaissance, et assure donc une indépendance théorique entre ses critères et ceux liés au système (calcul sur les taux de confiance, etc ...).

Le procédé retenu est de faire analyser la chaîne de candidats-phonèmes constituant le candidat-mot par un automate, qui décidera s'il faut ou non essayer de trouver ce mot dans le dictionnaire. L'ensemble des suites de phonèmes acceptées par l'automate est un large sur-ensemble du vocabulaire français, mais il réduit cependant considérablement le nombre de chaînes de phonèmes à chercher dans le dictionnaire.

Limiter cependant de la sorte les possibilités d'enchaînement des phonèmes revient à faire rejeter dans certains cas par l'automate des mots français, sous peine de sophistiquer inutilement ses règles de production. S'il existe de tels mots, indispensables au vocabulaire du système, on peut envisager de les regrouper dans un sous-dictionnaire d'exceptions, à accès particulier. Il s'agit donc, en résumé, de trouver un modèle qui, en quelque sorte, optimise les trois critères suivants :

- *puissance de sélection de l'automate*
- *simplicité de sa mise en oeuvre*
- *"couverture" maximale des mots du français*

2) Choix de l'automate

L'examen des enchaînements possibles des phonèmes dans la langue française a été réalisé dans une optique statistique par Haton [1], Rossi [2], Liénard [3]. On peut extraire des tableaux de diphonèmes qu'ils donnent, ainsi que des études de Juban [4] sur la syllabe française, et naturellement, d'un dictionnaire de la langue, une série de règles sur les enchaînements possibles de phonèmes à l'intérieur des mots du français. Ces constantes

proviennent de contraintes articulatoires, de règles d'économie, de règles diachroniques d'élocution, etc .... On peut considérer comme hypothèse de construction du modèle, que les contraintes sont de nature syllabique, c'est-à-dire que, si l'on note GC un groupe consonnantique (suite d'une ou plusieurs consonnes) dans un mot, GV un groupe vocalique, SS une semi-voyelle, on n'interdira dans un mot que la suite SS - GC ; et que les contraintes à l'intérieur d'un groupe n'influeront pas sur les autres groupes du mot.

Cette hypothèse sur le "contexte" nous a donc conduit à choisir un automate fini, fonctionnant de manière déterministe, comme modèle des contraintes phonologiques. Nous avons naturellement imposé qu'un mot ne puisse pas être entièrement composé de consonnes ; et d'autre part qu'il ne puisse ni commencer, ni finir par une semi-voyelle (ce qui n'élimine que très peu de mot "français").

La recherche des classes de phonèmes qui seront les terminaux de cet automate nous a conduit au choix indiqué à la figure 1. Il faut noter la cohérence de ces classes avec les définitions phonétiques traditionnelles, du moins en ce qui concerne les consonnes, ce qui rend compte des contraintes dans les enchaînements de mouvements articulatoires.

Les enchaînements possibles à l'intérieur des groupes vocaliques et consonnantiques sont indiqués en figure 2. Le modèle rejette pour le moment les triconsonnes, ce qui élimine principalement les groupes /s t r/ et /liquide - plosive voisée - liquide/. Les diconsonnes /g z/ ("x" doux) et /k t/ sont également éliminées. Mais d'une façon générale, ce modèle n'élimine que des groupements peu productifs.

La description formelle de ces règles nous a conduit à un automate fini possédant 13 états, non compris l'état final. Il fonctionne de manière déterministe. Un test lui a été rajouté afin de rendre compte des règles spéciales aux débuts de mots (cf figure 2).

Un calcul théorique montre que cet automate n'accepte qu'une sur six chaînes de trois phonèmes composée de façon aléatoire ; pour cinq phonèmes, la proportion passe à plus de 10. La réduction théorique est donc assez importante.

3) Mise en oeuvre

L'automate conduira donc la composition des candidats-mots à partir des candidats-phonèmes proposés. Le tableau est d'abord transformé en tableau de terminaux de l'automate. Celui-ci ne laissera construire, dans un parcours de ce tableau, que les suites de terminaux qu'il reconnaît. Celles-ci seront donc disponibles pour une vérification lexicale, après reconversion en leur forme originale de suite de phonèmes.
Un exemple de la mise en oeuvre est donné en figure 3. On peut apporter d'autre part une aide à cette prédiction, en utilisant à nouveau les taux de confiance affectés aux candidats-phonèmes : un calcul de taux de confiance global du candidat-mot peut être mené en parallèle avec sa construction et son analyse. Des seuils dans ce taux peuvent conduire à des rejets d'un autre type.
On peut envisager de ranger les candidats-phonèmes ligne par ligne, dans un ordre "alphabétique". Une conduite adaptée de l'analyse du tableau par l'automate assure que les candidats sélectionnés se présentent rangés dans cet ordre, ce qui diminue le nombre de recherches dans le dictionnaire.
Enfin, si des hypothèses de mauvaise segmentation du mot prononcé sont induites , par exemple, par l'échec de la recherche lexicale, l'analyse peut s'effectuer en supposant des omissions ou des insertions entre les lignes du tableau, qui représentent un segment du mot. L'automate fonctionnera ainsi pour prédire des mots sur un tableau corrigé.

4) Résultats

Des expériences ont été faites sur un ensemble de 200 mots, pour ceux de plus de deux phonèmes. D'autre part, le vocabulaire terminal ne comporte pour l'instant que 22 phonèmes : les plosives et les fricatives non voisées n'ont pas été envisagées. Il est donc difficile de faire une évaluation correcte sur ces données. Cependant, nous avons obtenu une moyenne de réduction légèrement inférieure aux taux théoriques cités plus haut, ce qui tend à montrer l'indépendance relative entre les critères de reconnaissance phonémique et ceux d'enchaînement phonologique.

Une telle méthode semble devoir s'appliquer à de gros dictionnaires, où le temps de recherche d'un mot est important, à cause non seulement du parcours de toutes les entrées, mais aussi du transfert éventuel de ce dictionnaire d'une mémoire de masse.
Le temps d'analyse d'un mot à reconnaître dépend naturellement du nombre de candidats-phonèmes de chacun de ses segments. Il ne semble cependant pas être assez important, compte-tenu du gain de recherche dans le dictionnaire, pour grever lourdement les performances du système de reconnaissance.

Conclusion

La méthode présentée propose donc une procédure syntaxique de prédiction de chaînes de phonèmes, vis-à-vis du vocabulaire français. Elle s'inscrit dans le modèle général "prédiction-vérification" qui semble s'imposer à tous les niveaux d'un système de reconnaissance évolué de la parole, fonctionnant de façon analytique et disposant d'un large vocabulaire.

Figure 1

Terminaux de l'automate

C 1 : Plosives non voisées /P/, /T/, /K/
C 2 : Plosives voisées /B/, /D/, /G/
C 3 : Liquides /L/, /R/
C 4 : Fricatives "labio-dentales" /F/, /V/
C 5 : /S/
C 6 : Nasales , autres fricatives /N/, /M/, /Z/, /Ʒ/, /ʃ/
S S : Semi-voyelles /J/,/W/, /ɥ/
V 1 : Voyelles équivalentes aux semi-voyelles /I/, /Y/, /OU/
V 2 : /E/
V 3 : Autres voyelles /A/, /O/, /OO/
/EE/, /AI/, /OE/
/AN/, /ON/, /IN/

Figure 2

Groupes consonnantiques acceptés

GC :	*C 1*	*C 3*	*"âpre"*
	C 2	*C 3*	*"vibrer"*
	C 5	*C 1*	*"casque"*
	C 1	*C 5*	*"axe"*
	C 4	*C 3*	*"frein"*
interdits en début de mot	*C 3*	*C 1*	*"halte"*
	C 3	*C 2*	*"argument"*
	C 3	*C 3*	*"parler"*
	C 3	*C 4*	*"élevage"*
	C 3	*C 5*	*"nerveux"*
	C 3	*C 6*	*"berge"*

Groupes vocaliques acceptés

G V :	*V 1*	*V 2*	*"buée"*
	V 2	*V 3*	*"néant"*
	V 1	*V 3*	*"boueux"*
	V 2	*V 1*	*"pays"*

Figure 3 - *Exemple de prédiction*

```
*****  \  I  L  ϴN   *****
```

Mot à reconnaître

```
***  \#  \#
***  I#  J#
***  L#  Y#  ϴE#  EE#  M#  V#  J#  ϴN#
***  ϴN#  EE#  ϴ#  #
```

Résultat de la reconnaissance phonémique : tableau des candidats-phonèmes. Nombre de candidats-mots à ce stade : 128

```
M#  \#
I#  J#
EE#  J#  L#  M#  ϴE#  ϴN#  V#  Y#
EE#  ϴ#  ϴN#  #
```

Tri du tableau par ordre alphabétique

```
C6      C6
V1      SS
V3      SS      C3      C6      V3      V3      C4      V1
V3      V3      V3      SS
IZ
```

Tableau des terminaux de l'automate correspondant au tableau trié de candidats-phonèmes

```
M#  I#  J#  EE#
M#  I#  J#  ϴ#
M#  I#  J#  ϴN#
M#  I#  L#  EE#
M#  I#  L#  ϴ#
M#  I#  L#  ϴN#
M#  I#  M#  EE#
M#  I#  M#  ϴ#
M#  I#  M#  ϴN#
M#  I#  V#  EE#
M#  I#  V#  ϴ#
M#  I#  V#  ϴN#
M#  J#  Y#  EE#
M#  J#  Y#  ϴ#
M#  J#  Y#  ϴN#
\#  I#  J#  EE#
\#  I#  J#  ϴ#
\#  I#  J#  ϴN#
\#  I#  L#  EE#
\#  I#  L#  ϴ#
\#  I#  L#  ϴN#
\#  I#  M#  EE#
\#  I#  M#  ϴ#
\#  I#  M#  ϴN#
\#  I#  V#  EE#
\#  I#  V#  ϴ#
\#  I#  V#  ϴN#
\#  J#  Y#  EE#
\#  J#  Y#  ϴ#
\#  J#  Y#  ϴN#
```

30 candidats-mots acceptés par l'automate, rangés par ordre alphabétique.

Aucun calcul de taux de confiance global n'est fait ici pour ces candidats.

```
IAP = 4.26667
```

Gain de recherches lexicales pour le mot à reconnaître

BIBLIOGRAPHIE

[1] *J. P. HATON "Contributions à l'analyse , la paramétrisation et la reconnaissance automatique de la parole" - Thèse d'Etat - 1974 - NANCY*

[2] *M. ROSSI "Au sujet des groupes consonnantiques du français" - Institut de phonétique - Aix-en-Provence*

[3] *J. S. LIENARD "Analyse, Synthèse et reconnaissance automatique de la parole" Thèse d'Etat - Paris VI - 1972*

[4] *P. JUBAN "La syllabe en français" - C N E T - Mars 1973*

[5] *A. MAISSIS "Le traitement de l'information acoustique, étape fondamentale de la reconnaissance de la parole" Thèse d'Etat - Paris VI - 1973*

[6] *R. J. SHOLES "Phonotactic grammaticality" - Mouton & Co - 1966*

[7] *C.R.E.D.I.F. "Dictionnaire du français fondamental - Publications de l'I.N.R.D.P.*

RECONNAISSANCE DE GRANDS DICTIONNAIRES PRONONCES PAR PLUSIEURS LOCUTEURS

R. VIVES, L. BUISSON, J-Y. GRESSER
G. MERCIER, M. QUERRE, C.N.E.T.-LANNION

Résumé.

Nous présentons une expérience de reconnaissance de mots isolés prononcés par plusieurs locuteurs. Deux approches principales sont testées : les résultats obtenus par la méthode analytique sont moins bons que ceux fournis par la méthode globale.
L'accent est mis sur les contraintes linguistiques introduites aux niveaux acoustique, phonétique et lexical.

Abstract.

Two word recognition systems were tested with several speakers. Better results were obtained with the "global" recognizer (using dynamic programming direct matching of acoustic patterns) than with the "analytical" recognizer (working on "phonetic" transcription).
Emphasis is put upon linguistic constraints which are introduced at the acoustic, phonetic and lexical levels.

RECONNAISSANCE DE GRANDS DICTIONNAIRES PRONONCES PAR PLUSIEURS LOCUTEURS

I. - INTRODUCTION.

Dans le contexte de la reconnaissance de grands dictionnaires prononcés par plusieurs locuteurs, nous nous proposons d'exposer les critères linguistiques qui nous ont paru intéressants dans la mise en oeuvre d'une approche analytique [1], [2], d'un système de communication homme-machine : KEAL (1) = Kenreizhadur (2), Evit (3), Anavezout (4), Lavar (5).

En entendant par critère linguistique tout critère fondé sur une formalisation plus ou moins profonde de la donnée acoustique, nous développerons plus particulièrement les aspects concernant les niveaux acoustique, phonétique et lexical.

Une comparaison entre la méthode analytique et une approche globale [3], [4], [5] de reconnaissance automatique de mots isolés est effectuée sur le plan des performances des machines.

II. - APPLICATION DE CRITERES LINGUISTIQUES DANS KEAL.

a) Niveau acoustique.

La parole d'entrée, échantillonnée à l'aide d'un vocodeur à canaux se présente au début du traitement sous la forme d'une suite de vecteurs à 15 dimensions (14 valeurs d'énergie recueillies dans les filtres, plus le pitch). La séparation bruit-parole dans ce signal est loin d'être triviale : nous nous contentons de tests rudimentaires comme la vérification du niveau d'énergie dans certains filtres et un comptage des échantillons présumés de parole. Il va de soi que de tels tests laisseront passer comme parole, tous les bruits de bouche prolongés que l'on peut, par exemple, proférer devant un micro.

Pour adapter nos machines à l'homme il sera intéressant de savoir dans quelle mesure on pourra discriminer la parole d'un raclement de gorge, sans aller pour autant jusqu'à la reconnaissance ou la compréhension complète du signal.

en breton : (1) idée,
(2) système,
(3) pour,
(4) connaître,
(5) parole.

La première étape de Keal inspirée de la linguistique est la segmentation de la forme acoustique en syllabes. Cette procédure repère dans la syllabe la position de la voyelle. Le segment voyelle ne sera comparé, par la suite, qu'à des éléments "voyelle" de référence et il ne sera plus possible à ce niveau de confondre des phonèmes comme par exemple m et i dont les réalisations acoustiques sont cependant très voisines [6].

b) Niveau phonétique.

Entre deux segments "voyelle", Keal recherche de nouveaux segments qu'il tente d'étiqueter à l'aide de phonèmes consonnantiques ou semi-consonnantiques.

1) Etiquetage des segments consonnes.

Keal utilise une procédure d'identification hiérarchisée (figure 1a) inspirée par une hiérarchisation linguistique idéale (figure 1b). Les tests de voisement ou de non-voisement (niveau i) ne sont effectués que dans le cas d'un doute au niveau de la séparation plosive/non plosive (niveau ii). En parallèle, Keal procède à une approche globale d'identification des segments qui fournit pour chaque segment une liste de phonèmes. Dans chaque liste, seuls sont retenus ceux des phonèmes qui définissent le noeud de la hiérarchie où l'on est arrivé (figure 2).

Dans le cas de segments en fin de mot, Keal supprime de la liste restante les phonèmes w ou y : il n'existe pas en français de mot se terminant par ces deux semi-voyelles.

2) L'étiquetage des segments "voyelle" est effectué par une méthode globale. Comme pour les consonnes nous envisageons d'employer un schéma d'identification hiérarchique fondé sur l'aperture et le lieu d'articulation des voyelles.

c) Niveau lexical.

C'est parce que le module de recherche lexicale de Keal utilise comme référence des dictionnaires de mots codés phonétiquement que l'on peut voir à ce niveau une application des contraintes de succession des différents phonèmes dans les mots.

Le module de segmentation et de détection de phonèmes donne du signal d'entrée une description formée d'une suite de réponses multiples de phonèmes munis d'un degré de confiance.

L'exemple suivant donne une forme de la description obtenue pour le mot "argent" :

p 0,3	a 0,9	w 0,6	ʒ 0,7	ã 0,8
t 0,3	ɔ̃ 0,8	ʁ 0,6	f 0,6	ʒ 0,8
k 0,3	ɛ̃ 0,8			ɸ 0,7
b 0,3				

Que cela soit par une méthode de décodage séquentiel ou par l'emploi d'un indice de ressemblance [3], la philosophie de la procédure de recherche lexicale est de réduire l'information contenue dans la suite des réponses multiples, dans le contexte du dictionnaire.

III. - RESULTATS DES EXPERIENCES ET CONCLUSION.

La reconnaissance de mots isolés par programmation dynamique, testée sur un corpus de plusieurs centaines de mots prononcés par plusieurs locuteurs, a fourni jusqu'à présent des résultats plus satisfaisants que par l'approche analytique. Nous en avons tiré deux enseignements :

1) Il nous a paru nécessaire de "globaliser" la méthode analytique : nous avons commencé par l'introduction d'un degré de confiance sur chaque segment trouvé. Dans les mots commençant par une voyelle le module de segmentation détecte "avec une certaine probabilité" un segment plosif. Considéré comme sûr au niveau de la recherche lexicale, ce segment supplémentaire avantageait les mots plus longs que le mot prononcé. Le mot prononcé "elle" (ɛl) est reconnu plus facilement quand on tient compte du degré de confiance sur les segments, alors que "belle" (bɛl), "pelle" (pɛl) ou "quel" (kɛl) sont reconnus à sa place dans l'autre cas.

Nous envisageons, d'autre part, d'employer directement la programmation dynamique pour une reconnaissance phonétique plus fine.

2) Les critères linguistiques utilisés dans la méthode analytique semblent avoir une importance non négligeable sur les performances de Keal. Il ne serait pas étonnant qu'ils eussent des répercussions intéressantes s'ils étaient appliqués à des méthodes globales de reconnaissance de la parole. Nous pensons surtout à résoudre de cette façon les problèmes d'accès à une grande masse de données.

BIBLIOGRAPHIE

[1] J-Y. GRESSER, G. MERCIER - Automatic segmentation of speech into syllabic and phonemic units. Application to french words and utterances.
Symposium on "Auditory analysis and perception of speech", Leningrad, August 1973.

[2] L. BUISSON, G. MERCIER, J-Y. GRESSER, R. VIVES, M. QUERRE - Uncertainties of objective phonetic segmentation. Model and applications.
Eighth international congress on acoustic, London 1974.

[3] R. VIVES, J-Y. GRESSER - A similarity index between strings of symbols. Application to automatic word and language recognition.
First international joint conference on pattern recognition, Washington, D.C., October 30 - November 1, 1973.

[4] J-P. HATON - Contribution à l'analyse, la paramétrisation et la reconnaissance automatique de la parole - Thèse d'Etat, Nancy 1974.

[5] M. QUERRE, G. MERCIER, J-Y. GRESSER - Reconnaissance automatique de la parole : application de la programmation dynamique à l'identification de mots isolés. Résultats comparés sur 1000 mots. Note Technique Interne 1973.

[6] M. CARTIER - Reconnaissance de vocabulaires spéciaux.
Note Technique Interne TMA/ETA/23, 1974, pp. 38-45.

Identification des consonnes et semi-voyelles.

fig 1a

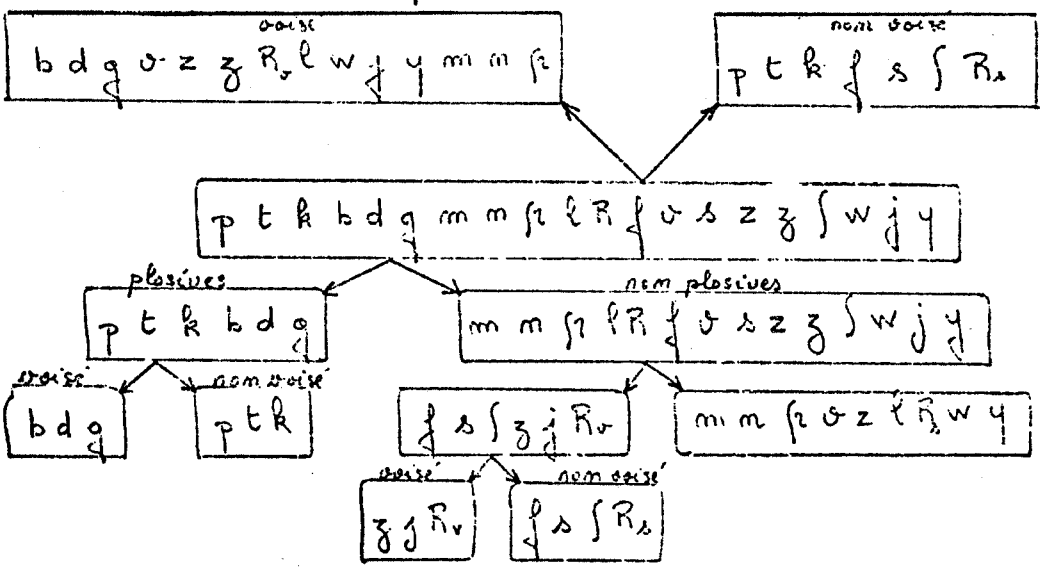
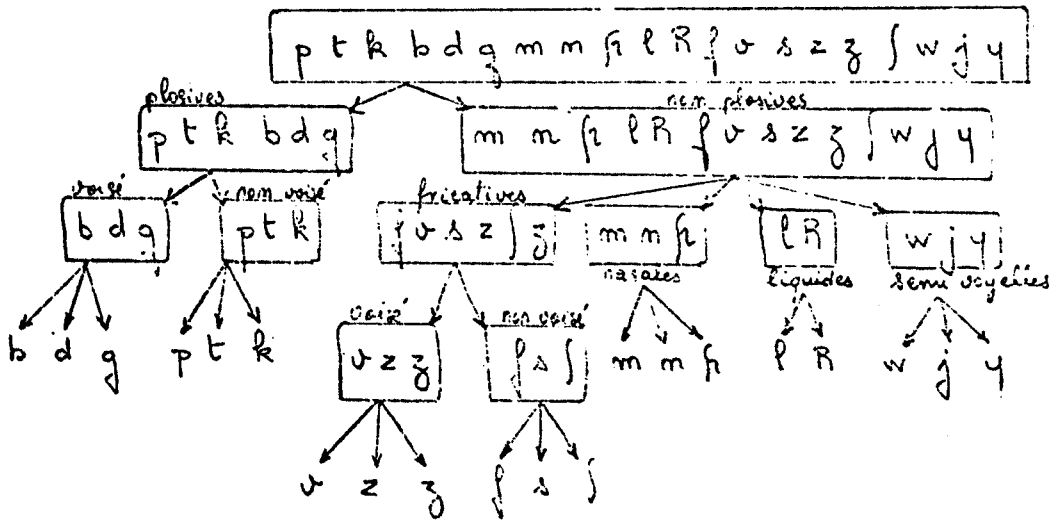


fig 1b



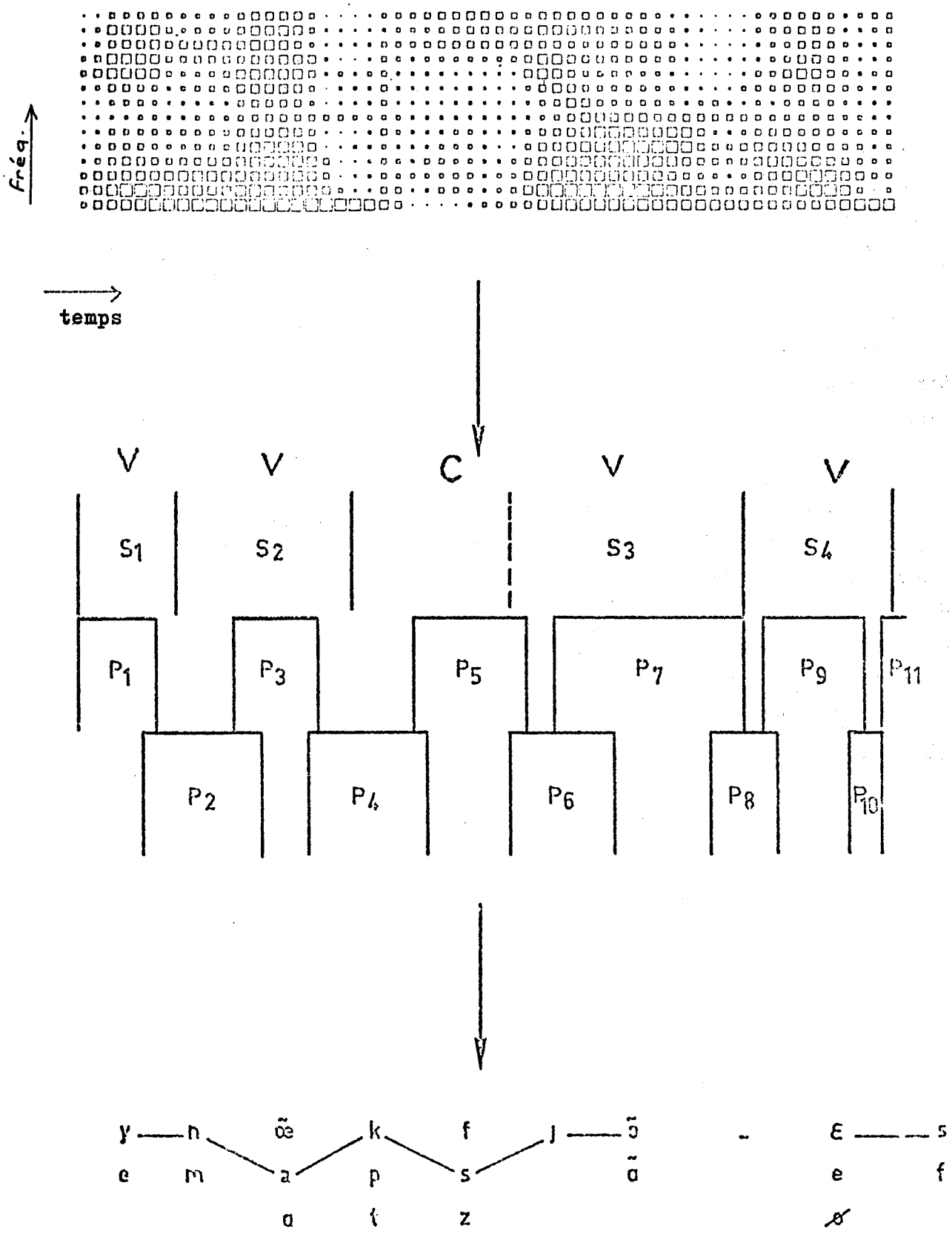


Fig. 2 - Reconnaissance phonétique à partir de Keal

Reconnaissance subjective et objective de la parole codée (phonocode)

par J.A. Dreyfus-Graf et coll., Genève, et CNET, Lannion.

Résumé

Les codes phonétiques, nommés "phonocodes" sont des langages logiques destinés à la commande verbale de machines, en temps réel. Ils sont basés sur des nombres restreints de classes de phonèmes, tels que O,I,A,S,T,N pour le code SOTINA, ou O,I,A,Ŝ,K,N pour le code ŜOKINA. Des tests subjectifs ont montré que ces deux codes permettent d'obtenir des taux d'erreur moyens de un-pour-mille par phonème, et qui sont du même ordre que ceux d'une bonne dactylographe. Pour les langues naturelles, les taux correspondants sont de l'ordre de dix-pour-cent, soit cent fois plus élevés. Les tests objectifs seront effectués à l'aide de machines de reconnaissance de la parole codée telles que le "phono-décodeur I", qui est en cours de construction.

Summary

The phonetic codes referred to as "phonocodes" are logical languages intended for the verbal operation of machines in real time. They are based on restricted numbers of phoneme classes, such as O,I,A,S,T,N for the SOTINA code, or O,I,A,Ŝ,K,N for the ŜOKINA code. Subjective tests have shown that these two codes admit average error-rates of one-per-thousand per phoneme, which are of the same order as the error-rates of good typists. For natural languages the corresponding error-rates are of the order of ten-per-cent, i.e. hundred times larger. Objective tests will be made with the help of recognition machines, such as the "phono-decoder I" which is in course of construction.

Reconnaissance subjective et objective de la parole codée (phonocode)

par J.A. Dreyfus-Graf et Coll., Genève, et CNET, Lannion.

1. Buts des phonocodes

Une langue naturelle, telle que le français ou l'anglais, comporte quelque 32 phonèmes, dont les simples arrangements, à 4 phonèmes chacun, représentent déjà 1 million de mots possibles ($32^4 = 2^{20} \simeq 10^6$). Il est donc évident qu'une langue naturelle n'est pas du tout adaptée à la commande verbale d'appareils, quand il s'agit de vocabulaires limités à quelques dizaines ou centaines de mots.

Il semble alors préférable de développer des langages logiques et simplifiés, qu'on peut nommer "phonocodes" [1,2,3,4,5,6] . Ils doivent permettre de commander, internationalement et en temps réel, des appareils, tels que sélecteurs téléphoniques, trieuses postales, télé-imprimeurs, serrures ou calculateurs. Les machines reconnaissant la parole codée, et qu'on peut nommer "phono-décodeurs" n'auront pas besoin de mémoriser préalablement chaque vocabulaire, voire chaque mot, prononcé par chaque locuteur, mais se trouveront immédiatement adaptées à l'ensemble des usagers.

2. Choix des classes de phonèmes

Contrairement aux langues naturelles qui exigent plus de 30 phonèmes, les phonocodes se contentent d'un nombre de classes de phonèmes inférieur à 15, et qu'on peut choisir parmi les suivantes :

Fig.1

	voyelles						consonnes								
numéro hiérarchique	1.	2.	3.	4.	5.	6.	1.	2.	3.	4.	5.	6.	7.	8.	9.
symbole graphique	O	I	A	E	U	Y	T	S	N	Ŝ	P	R	L	D	Z
variante	U	E		Ê			K	Ŝ	M	H					J
valeur phonétique	o	i	a	e	u	y	t	s	n	ʃ	p	r	l	d	ʒ
variante	u	e	ɑ	ɛ			k	ʃ	m						

Le nom de chaque phonocode est formé par une association des consonnes et voyelles qui le composent :

Fig.2

nom	consonnes	nombre	voyelles	nombre	total phonèmes
SOTINA	S,T,N,	3	O,I,A	3	6
ŜOKINA	Ŝ,K,N	3	O,I,A	3	6
SOTINAKEMUŜ	S,T,N,K,M,Ŝ	6	O,I,A,E,U	5	11
SOTINAKEMUŜYR	S,T,N,K,M,Ŝ,R	7	O,I,A,E,U,Y	6	13

Pour déterminer les sons et symboles qui fournissent les taux d'erreur minimums, c'est-à-dire le maximum de fiabilité, on procédera à des tests subjectifs et objectifs, d'une part à l'aide de locuteurs et d'auditeurs humains, d'autre part avec le concours de machines qui reconnaissent et qui synthétisent la parole automatiquement.

3. Tests subjectifs avec le code SOTINAKEMUŜ

Les matrices d'erreur de la Fig.3 résument 5 séries de tests subjectifs N° 1 à 5 effectués avec le SOTINAKEMUŜ. Celui-ci englobe les 6 consonnes S,T,N,K,M,Ŝ, et les 5 voyelles O,I,A,E,U.

Test 1. Liaison directe SFERT, micro dynamique, sans filtre, sans bruit superposé

ENTENDU (erreurs en "pour-cents" = % = 10^{-2}) \ PRONONCÉ	K	T	Ŝ	S	M	N	A	I	E	O	U	ϕ	TOTAL PAR CODE	MOYENNE " (%)
TOTAL PAR PHONÈME	2,4	0	1,33	3	2,1	0,4	0,4	0,3	0	0	1,7		= 10,68/11	= 1 %
∀			0,28											
ϕ	0,4						0,4							
U														
O											1,7			
E								0,3						
I							0,4							
A														
N					1,6									
M				0,3										
S			0,8		0,5									
Ŝ														
T	2		0,3											
K														
SOTINA		0		0		0	0,8	0		0		0	0,8	0,13 %
ŜOKINA	0,4		0			0	0,8	0		0		0	1,2	0,2 %

MOYENNE SOTINAKEMUŜ

Test 2. Liaison téléphon. S.63, micro magnétique, 300-3400 Hz, bruit salle 60 dB (S/B= +6dB)

ENTENDU (erreurs en "pour-cents" = % = 10^{-2}) \ PRONONCÉ	K	T	Ŝ	S	M	N	A	I	E	O	U	ϕ	TOTAL PAR CODE	MOYENNE " (%)
TOTAL PAR PHONÈME	6	4,5	1,2	3,1	2	3,3	0,4	0,4	1,1	0	0	0,1	= 20,6/11	= 2 %
∀	1	0,5		0,6		0,3						0,1		
ϕ		1,5		0,6	0,5									
U														
O							0,4							
E								0,4						
I									0,7					
A														
N														
M						1,8								
S			1,2											
Ŝ				1					0,4					
T	5				1	1,4								
K		2,5		0,9	0,5	0,3								
SOTINA		1,5		1,5		1,4	0,4	0		0		0,1	4,9	0,8 %
ŜOKINA	0		0			0,3	0,4	0		0		0,1	0,8	0,13 %

MOYENNE SOTINAKEMUŜ

Test 3. Liaison téléphon. U.43, micro à charbon, 300-3400 Hz, bruit salle 60 dB (S/B= +6dB)

ENTENDU (erreurs en "pour-cents" = % = 10^{-2}) \ PRONONCÉ	K	T	Ŝ	S	M	N	A	I	E	O	U	ϕ	TOTAL PAR CODE	MOYENNE " (%)
TOTAL PAR PHONÈME	2,0	3,3	2,1	5,7	1,9,1	2	0,3	2	5,2	2	2,3	0,3	= 79,5/11	= 7 %
∀		1,8		0,6		0,4		0,4				0,3		
ϕ		1		0,3		0,4			0,8		0,3			
U	0,5					0,4		0,4	0,4	0,8				
O											2			
E		0,5					0,3	1,2						
I	0,5								2,1					
A									1,2	1,8				
N				0,5	10					0,4				
M		0,5												
S			2		0,4	0,4								
Ŝ		0,5		4										
T	1,0					0,4			0,4					
K		5		0,3										
SOTINA		1,5		1,7		2	0	0,4		2		0,3	7,9	1,3 %
ŜOKINA	1		0			1,6	0	0,1		2		0,3	5,3	0,9 %

MOYENNE SOTINAKEMUŜ

Test 4. Liaison directe SFERT, micro dynamique, sans filtre, bruit 72 dB (S/B = -6)

ENTENDU (erreurs en "pour-cents" = % = 10^{-2}) \ PRONONCÉ	K	T	Ŝ	S	M	N	A	I	E	O	U	ϕ	TOTAL PAR CODE	MOYENNE " (%)
TOTAL PAR PHONÈME	11,7	2,1	2,2	5	5,8	5,1	0	0,6	1,3	2,5	0,8	1,3	= 39,6/11	= 3,6 %
∀														
ϕ	4,0	0,5	1,1	1		0,4						0,4		
U					0,4			0,3		2,5				
O														
E														
I									1,3		0,4			
A														
N	0,8				3,7			0,3				0,1		
M						1,7						0,1		
S		0,5	1,1									0,2		
Ŝ				1,3										
T	7			1,7	0,4	0,3						0,3		
K		1,1		1	1,3							0,1		
SOTINA		1		3,7		3,4	0	0,3		0		0,8	9,2	1,5 %
ŜOKINA	5,7		1,1			0,4	0	0,3		0		0,8	8,3	1,4 %

MOYENNE SOTINAKEMUŜ

Test 5. Liaison directe SFERT, micro dynamique, 300-3400 Hz, sans bruit superposé

ENTENDU (erreurs en "pour-cents" = % = 10^{-2}) \ PRONONCÉ	K	T	Ŝ	S	M	N	A	I	E	O	U	ϕ	TOTAL PAR CODE	MOYENNE " (%)
TOTAL PAR PHONÈME	1,2	0,7	2,5	3,8	1,1	0,5	0	0	0	0	0	0,1	= 9,9/11	= 0,9 %
∀				0,8										
ϕ						0,5								
U														
O														
E														
I														
A														
N					1,1									
M														
S			2,5											
Ŝ				3										
T	1,2											0,1		
K		0,7												
SOTINA		0		0		0,5	0	0		0		0,1	0,6	0,1 %
ŜOKINA	0		0			0,5	0	0		0		0,1	0,6	0,1 %

MOYENNE SOTINAKEMUŜ

Fig. 3 TESTS SUBJECTIFS AVEC LES PHONOCODES SOTINAKEMUŜ, SOTINA, ŜOKINA

Matrice des taux d'erreur (%), par phonème.

Niveau de parole = 66 dB
Nombre maximum de classes de voyelles V = 5
" " " " " consonnes C = 6
" " " " " phonèmes = 11

Nombres de phonèmes par "codatome" = 3
("codatome = logatome codé)
Configurations : VCV, CVC, CVV, CCV, VCC

Nombre de codatomes émis, par test : ~ 250
" " " reçus, " " : ~1000
Nombre de phonèmes par colonne : ~ 280
" " " " code : ~3080 ou ~1680

Taux d'erreur moyen = somme des taux des classes concernées, divisée par le nombre de ces classes, c'est-à-dire 11 pour le code SOTINAKEMUŜ, 6 pour le code SOTINA ou ŜOKINA

ϕ = phonèmes manqués; ∀ = phonèmes inventés
Ŝ = "ch" = /ʃ/ ; E = /e/ ; U = "ou" = /u/ ; O = "ô" = /o/

Concernant le code SOTINAKEMUŜ , et avec un niveau de parole de 66 dB, les taux d'erreur moyens par phonème ont été les suivants :

1% : avec microphone dynamique, sans filtre et sans bruit
0,9% : dans les mêmes conditions, mais avec filtre téléphonique (300-3400 Hz)
2% : avec microphone magnétique (S.63), filtre téléphonique et bruit normal de 60 dB
3,6% : avec microphone dynamique, sans filtre, mais avec bruit exagéré de 72 dB
7% : sans bruit, mais avec microphone téléphonique à charbon (U.43).

On constate que la présence ou l'absence d'un filtre téléphonique (passe-bande 300 à 3400 Hz) ne change pratiquement rien. D'autre part, un bruit exagéré, dépassant même de 6 dB le niveau de la parole, est pourtant moitié moins nocif que la simple intervention d'un microphone à charbon.

4. Tests subjectifs avec les codes SOTINA et ŜOKINA

Les taux d'erreur des codes SOTINA et ŜOKINA sont déduits de ceux du code SOTINAKEMUŜ qui figurent dans la Fig.3 en ne tenant compte que des phonèmes concernés.

La Fig.4 récapitule les divers taux d'erreur moyens, par phonèmes, correspondant aux 3 codes mentionnés, et elle les compare aux taux théoriques, prédits lors des 4èmes Journées d'Etude sur la Parole [7] . Les ordres de grandeurs des valeurs mesurées sont bien en accord avec ceux des valeurs calculées.

			TAUX D'ERREUR			subjectifs (mesurés)				théoriques*		
			avec nombres de phonèmes →			11	6	6	32	32	11	6
Test No	liaison	micro	filtre Hz	bruit dB	rapport S/B	SOTINA -KEMUŜ	(codatomes) SOTINA	ŜOKINA	logatomes FRANÇAIS*			
1.	directe	dynam.	sans	sans	+66	1 %	0,13%	0,2 %	10 %	10%	1%	0,1%
2.	téléph,	S.63	300-3400	60	+6	2 %	0,8 %	0,13%	20 %	20%	2%	0,2%
3.	téléph.	U.43	300-3400	60	+6	7 %	1,3 %	0,9 %	30 %	30%	3%	0,3%
4.	directe	dynam.	sans	72	-6	3,6%	1,5 %	1,4 %	40 %	40%	4%	0,4%
5.	directe	dynam.	300-3400	sans	+66	0,9%	0,1 %	0,1 %	10 %	10%	1%	0,1%

Fig.4. Récapitulation des taux d'erreur moyens, par phonème : subjectifs (mesurés) et théoriques (calculés). Niveau normal de parole: 66 dB. *ordre de grandeur

Examinons d'abord des liaisons directes, sans bruit superposé :

Selon qu'il s'agit de 32, 11 ou 6 phonèmes, les taux d'erreur sont dans le rapport de 10%, 1% et 0,1%. Ainsi les codes à 6 phonèmes admettent seuls des taux d'erreur de l'ordre de un-pour-mille, similaires à ceux d'une bonne dactylographe, tandis que les logatomes de la parole naturelle, avec leurs 32 phonèmes, impliquent des taux d'erreur cent fois plus élevés.

Il semble donc que les entrées vocales ne pourront concurrencer les entrées manuelles qu'avec l'aide de phonocodes, si la phase d'apprentissage préliminaire, par la machine, doit être évitée.

Que se passe-t-il dans le cas d'une liaison téléphonique à microphone magnétique, et en présence d'un bruit normal de 60 dB ? Le code ŜOKINA maintient un taux d'erreur de 1,3 pour-mille, contrairement au code SOTINA , dont le taux d'erreur augmente à 8 pour-mille. Ainsi ŜOKINA est 6 fois meilleur que SOTINA , du point de vue subjectif, c'est-à-dire par rapport à l'oreille humaine.

5. Tests objectifs

Les conclusions définitives, concernant les codes les plus favorables, dépendront encore des tests objectifs, c'est-à-dire des machines de reconnaissance utilisées. Un phono-décodeur I est actuellement en construction. C'est une machine spécialement adaptée à la reconnaissance d'un code à 6 phonèmes, environ, tel que ŜOKINA ou SOTINA [7,8].

Elle comprend essentiellement une partie analogique qui extrait les traits distinctifs des 6 classes de phonèmes, puis deux parties logiques qui appliquent les règles phonétiques et linguistiques du code, pour segmenter ces classes et en distinguer les associations.

La partie analogique utilise des compresseurs d'amplitude qui normalisent les niveaux de la parole et qui séparent les phonèmes plosifs des non-plosifs.

[4] . Ensuite elle peut employer l'une des nombreuses techniques d'analyse spectrales connues [9] -.

Pour effectuer les tests objectifs du code ŜOKINA , on peut se baser sur les matrices de mots à 2 ou 3 phonèmes indiquées précédemment pour le code SOTINA [7], en remplaçant T par K et S par Ŝ . Pour commencer les phono-décodeurs mémorisent les classes internationales de phonèmes, mais sont amnésiques aux mots qui les combinent. Il est donc indifférent que les listes des mots de tests soïent aléatoires ou ordonnées.

Les étapes des tests objectifs seront les suivantes :
D'abord reconnaissance des classes de phonèmes à l'intérieur des mots, ensuite - seulement - reconnaissance des mots élémentaires et non-élémentaires, formés par les combinaisons de ces classes. L'approche est donc inverse de celle qui évalue d'abord les formes externes (external pattern recognition [10]). Les choix définitifs des classes de phonèmes, ainsi que des règles phonétiques et linguistiques du code, seront conditionnés par les taux d'erreur minimums.

Une communication complémentaire à ce sujet est prévue pour le prochain Congrès International d'Acoustique (8ème C.I.A., Londres, 26-31 juillet 1974).

Les mesures de tests figurant dans la présente communication ont été effectuées au CNET, Lannion, par le Département ETA , en collaboration avec l'équipe de téléphonométrie du Département PRL.

R é f é r e n c e s

[1] J.A. DREYFUS-GRAF — Machines obéissant à la parole (Speech Responsive Machines). Conférence 20-C-15. Budapest, 20 août 1971, Proceedings 7th International Congress on Acoustics (ICA).

[2] " " " — Speech Dynamics and Pitch, Speech Symposium, Szeged, 27 août 1971 The Present State and Future Tasks of Speech Research, Round Table

[3] " " " — Machines commandées par la parole (Phonétographe V). Conférence du 7 déc. 1971, Revue A.I.M. Institut Montefiore, Liège, n°3 , 1972.

[4] " " " — Parole codée (phonocode): reconnaissance automatique de langages naturels et artificiels, Conférence du 21 janv. 1972, CNAM, Paris, Revue d'Acoustique, N°21, 1972.

[5] " " " — Recognition of Natural and of Artificial Speech (Phonocode). Communication H 9 du 25 avril 1972. Newton-Boston. Reports of the International Conference on Speech Communication and Processing. IEEE-AFCRL.

[6] " " " — Reconnaissance automatique de la parole codée (Phonocode) sonore et chuchotée. Conférence du 20 oct. 1972, GALF, ORTF. Issy-les-Moulineaux, Revue d'Acoustique, Paris, N°25, 1973.

[7] " " " — Codes phonétiques (Phonocodes) et règles linguistiques, 4èmes Journées d'Etudes, G.A.L.F., Groupe Communication Parlée, Institut de Phonétique, Université Libre, Bruxelles, 1973 (à paraître dans les Rapports)

[8] " " " — Tests d'intelligibilité de la parole codée (Phonocodes), Symposium G.A.L.F., F.A.S.E., A.B.A.V., Liège 1973 (à paraître dans les Rapports)

[9] A. ICHIKAWA, Y. NAKANO, K. NAKATA, Evaluation of Various Parameter Sets in Spoken Digits Recognition, IEEE, Audio and Electroacoustics, June 1973, Vol AU-21, N°3.

[10] W.A. LEA, An Approach to Syntactic Recognition Without Phonemics, IEEE, Audio and Electroacoustics, June 1973,Vol. AU-21, N3

PROJET D'UN CLASSIFICATEUR ACOUSTIQUE CONTRÔLÉ PAR UNE SYNTAXE

R.DE MORI - E.PICCOLO - S.RIVOIRA - A.SERRA

Centro di Studio per l'elaborazione numerale
dei segnali (CNR)

Politecnico di Torino
Istituto Elettrotecnico Nazionale G.Ferraris

Résumé

On décrit brièvement un classificateur acoustique de noyaux syllabiques obtenus avec un procédé de segmentation contrôlé par une syntaxe.
Dans chaque noyau syllabique, les évolutions des formants sont décrites par un autre langage artificiel et pour chaque noyau une séquence hiérarchique d'hipothèses est émise.

Abstract

The paper describes the main features of an acoustic classifier of sillable nuclei delimited by a linguistic procedure.
For every nucleus, the formant evolutions are described by an artificial language and for each nucleus a set of possible phonetic transcriptions is produced.

Introduction

Récemment on a adressé beaucoup d'efforts au problème de la reconnaissance automatique du language parlé et de la réalisation de systèmes de compréhension du parlé.
On a atteint d'une manière générale l'accord, que ces systèmes doivent travailler sur de differents niveaux interconnexes de procédé, comme la segmentation et la transcription phonétique.
Bien qu'il ait été demontré qu'une segmentation phonétique parfaitement exacte est pratiquement impossibile, même si l'on emploi l'oreille humaine comme analyseur, des expériences préliminaires, avec des systèmes de compréhension du parlé, permettent d'arriver à la conclusion qu'un classificateur acoustique-phonétique pourrait permettre un développement significatif des capacités du système d'identification.
La méthode proposée, c'est-à-dire, le développement naturel d'autres travaux faits sur la reconnaissance automatique des mots isolés dits en Italien, a son début de l'hypothèse étayée des expériences que les sons sont organisés comme une succession d'états du conduit vocalic . Ces états ne sont pas individuellement finalisés pendant la prononciation d'une phrase, pourtant le mot prononcé montre sans cesse des effects

de coarticulations. Pour être efficace, un classificateur acoustique doit considerer, comme des ensembles, les parts des données acoustiques, où il peut-être que la coarticulation est prédominante. La segmentation sera contrôlée par l'entremise d'une syntaxe, qui emploi un langage de toutes les possibles coarticulations.
De plus, la classification phonétique sera vue comme une génération (possiblement hiérarchique) d'hypothétique séries de phonèmes pour chaque segment de coarticulation. Cette génération sera contrôlée par une autre syntaxe, en employant une grammaire, qui permet de tirer des descriptions de toutes les données acoustiques associées avec la prononciation d'une certaine séquence de phonèmes dans de différents contextes par de différents orateurs dans de différentes situations.
La grammaire doit être applicable à tous les possibiles listes de phonèmes.

1) Méthodes de calcul du spectre

Actuellement trois méthodes pour le calcul des spectres sont en fonction. La plus rapide consiste dans un banc de 24 filtres de Tchebicheff 1/3 octave dans le domaine de fréquence 100 Hz/10.000 Hz; les sorties de ce banc sont échantillonnées à une fréquence de 100 Hz et le spectre d'une phrase est mis en mémoire sur un disque en temps réel [5].
Une méthode plus soignée, mais plus lente, emploie le calcul d'une FFT synchrone avec le pitch et permet d'obtenir 256 échantillons du spectre dans l'intervalle 0 - 5 kHz [1].
Enfin, pour une meilleure détection des formants, on dispose d'un algorithme de filtrage linéaire inverse qui donne une estimation de très grande vraisemblance [4].
Les échantillons du signal sont pris à groupes de N, où N est un multiple du piteh obtenu par l'algorithme de Meo [1]
Quand l'algorithme ne trouve pas une valeur du piteh acceptable, N est égal à 128 échantillons.
Cette méthode d'identification se base sur l'hypothèse que le signal vo-

cal soit stationnaire par moments et avec une excitation à densitée spectrale constante. On à développé aussi un système d'identification itératif dont le modèle est variable à chaque échantillon, en pouvant avoir soit des poles, soit des zéros et en permettant la détection du pitch [6].

2) Description automatique des charactéristiques globales des spectres et d'autres paramètres secondaires

L'ampleur A et le pitch P du message acoustique sont décrits par un langage artificiel L1 généré par une grammaire G1. Egalement les évolutions de deux paramètres, RV et LV, qui représentent l'aspect global du spectre, sont décrites par un autre langage artificiel L2 généré par une grammaire G2. R_V est le rapport entre l'énergie de basses et de hautes fréquences; LV est un paramètre qui dépend seulement de l'énergie de hautes fréquences. Deux descriptions, une (LD1) dans le langage L1 et l'autre (LD2) dans le langage L2, sont obtenues après la prononciation d'une phrase. Ces deux descriptions sont analysées par un automate sous le contrôle d'une grammaire G3; la sortie de cet automate est une séquence de symboles, et chacun d'eux est suivi par une référence temporelle. Les symboles correspondent à quatre classes dans lesquelles les traits du signal vocal sont premièrement classifiés. Ces classes sont:

(UT) : traits sans pitch;
(V) : voyelles;
(VC) : traits de consonnes sonores;
(SL) : silences

3) Segmentation

L'algorithme de segmentation subdivise le signal acoustique en segments qui correspondent à plusieurs phonèmes, pour lesquels on présume d'y trouver des effects évidents de coarticulation, qui empêcheraient une classification correcte des phonèmes au dehors du contexte. Le segments pseudo-syllabiques (PSS) ainsi obtenus peuvent être partiellement superposés. La segmentation est effectuée par un automate qui travaille sous le con-

trôle de la grammaire suivante:

$$G4 : \{V_t, V_n, P, (PSS)\}$$

$V_t = \{ (SL), (V), (VC), (UT)\}$

$V_n = \{ (PSS), (VLK), (UN) \}$

$(VLK) \longrightarrow (V)/(VC)(V)$

$(UN) \longrightarrow (SL)/(UT)/(SL)(UT)(SL)$

$(PSS)(VLK) \longrightarrow (V)(VLK)/(UN)(VLK)(VLK)$

$(PSS)(UN) \longrightarrow (V)(VLK)(UN)/(UN)(VLK)(UN)/(V)(VLK)(VC)(UN)/(UN)(VLK)(VC)(UN)$

4) Extraction et description des formants dans les traits sonores

Les traits sonores sont délimités dans les segments syllabiques par un simple automate qui compose la chaine la plus longue dans un segment avec les fragments V e VC, selon la règle suivante:

$$VCT \longrightarrow \alpha V \beta$$

$$\alpha \longrightarrow V/VC/V(VC)/e$$

$$\beta \longrightarrow VC/e$$

Pour chaque spectre dans un segment VCT, les zones de haute concentration d'énergie sont mises en évidence et jointes dans la dimension temporelle, selon un critère de maximisation d'une fonction objective. Les chaines élémentaires ainsi obtenues sont réduites aux formants avec un algorithme de successives éliminations de chaines redondantes.

On arrive ainsi à des formes qui peuvent être décrites par un langage défini par l'expression régulière suivante:

$$(VCT) = (F1T + F2T + F3T)$$

où le trait FiT est l'ensemble de i formants. Chaque trait FiT est décrit par un langage déjà employé pour la reconnaissance automatique de mots isolés et certains opérateurs décrivent les possibles concaténations des lignes d'un trait.

5) Emission d'hypothèses pour chaque segment syllabique sous le contrôle d'une grammaire des formes du langage parlé

Pour permettre une analyse efficace des erreurs du système, on a concen-

tré l'attention sur un sous-langage applicable à la réservation automatique de pièces sur avions, trains, etc. Le sous-langage a un vocabulaire d'environ 100 mots qui peuvent être employés pour former un millier de phrases acceptables. En outre, le sous-langage employé peut être construit en commençant de 150 noyau syllabiques.

Chaque noyau syllabique a une grammaire associée qui contient les règles de génération de possibles, différentes réalisations spectrales du noyau même. Chaque grammaire contient aussi des productions qui sont valables seulement si l'on produit des conditions pour les attributs de certains éléments.

Ces conditions concernent spécialement les paramètres qui dépendent du locuteur, c'est-à-dire les positions de concentrations d'énergie des voyelles stables et des fricatives.

La grammaire de chaque noyau est modifiée par un procédé d'inferance grammaticale, toutes les fois qu'un noyau n'est pas reconnu correctement. Pourtant il s'ensuit que quand une phrase est prononcée, la réalisation spectrale de chaque noyau peut être reconnue par plusieurs grammaires et qu'une phrase prononcée est représentée, à la sortie du classificateur acoustique, par une matrice où une ligne contient les possibles transcriptions phonétiques d'un noyau syllabique.

Cette transcription peut devenir hiérarchique si l'on emploie des grammaire stochastiques.

Conclusions

La figure 1 résume la conception du classificateur acoustique. Une description plus détaillée se trouve en [8].

Ce travail rentre dans le programme de recherche du "Centro per l'Elaborazione Numerale dei Segnali" et a été financié par le "Consiglio Nazionale delle Ricerche".

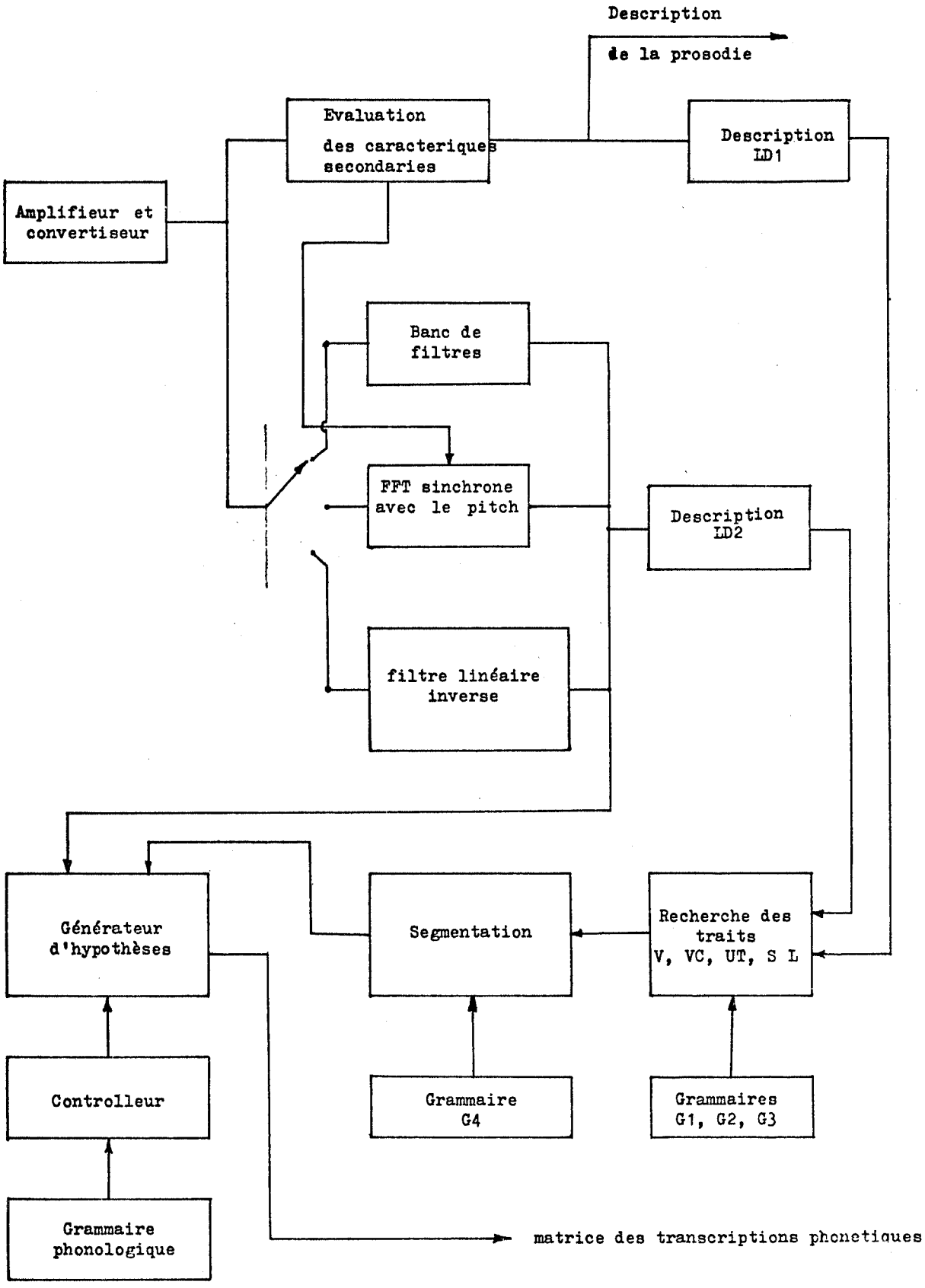

Description
de la prosodie
Evaluation
des caracteriques
secondaries
Description
LD1
Amplifieur et
convertiseur
Banc de
filtres
FFT sinchrone
avec le pitch
Description
LD2
filtre linéaire
inverse
Générateur
d'hypothèses
Segmentation
Recherche des
traits
V, VC, UT, S L
Controlleur
Grammaire
G4
Grammaires
G1, G2, G3
Grammaire
phonologique
matrice des transcriptions phonetiques

Fig. 1

Bibliographie

(1) M.Guglielmo, A.R.Meo, M.Mezzalama: "C.V.S.: a technique for generating a type of visible speech based on synchrons spectral analysis" (44th Audio Eng.Soc.Convention, Rotterdam, February 1973)

(2) R.De Mori: "Experiments in the automatic segmentation of continous speech" (44th Convention Audio Engineering Society, Copenhagen, 25-29 March, 1974)

(3) R.De Mori: "A descriptive technique for automatic speech recognition" (IEEE, Trans. vol. AU 21, n.2, 1973, p.89)

(4) E.Piccolo, A.Serra: "Extraction of acoustical correlates for automatic speech recognition using linear prediction " (8th Int.Congr. on Acoustics, London, 23-31 July, 1974)

(5) S.Rivoira, A.Serra: "Real-time automatic detection of syllable nuclei in continuous speech" (8th Int.Congr. on Acoustics, London, 23-31 July, 1974)

(6) A.Serra: "Identification of speech parameters using recursive identification techniques" (Speech Communications Seminar, Stockholm, August 1-3, 1974)

(7) E.Piccolo, A.Serra: "Pitch synchronons linear prediction analysis of speech wave for synthesis and recognition" (Technical report CENS, December 1973)

(8) R.De Mori: "Design for a syntax-controlled acoustic classifier" (IFIP Congress, Stockholm, August 1974)

COMMUNICATIONS LIBRES

INFLUENCE DE LA FREQUENCE FONDAMENTALE SUR L'ESPACE PERCEPTIF DES VOYELLES. *

R. BEECKMANS, R. CARRE et P. JOSPA.

Résumé.

Nous avons testé l'intelligibilité de sept voyelles du français, synthétisées à différents F_o tout en maintenant F_1 et F_2 constants (valeurs correspondant à une voix d'homme).

La baisse de performance obtenue pour les fréquences fondamentales élevées met en évidence un processus d'ajustement du système perceptif en fonction de F_o.

Une analyse des données par INDSCAL suggère quelques hypothèses concernant ce mécanisme.

Summary.

The intelligibility of seven french vowels, synthetised for different F_o keeping F_1 and F_2 constant has been tested. The drop in performance obtained for higher fundamental frequencies reveals an adjustment processed by the perceptual system in function of F_o.

An INDSCAL analysis of the data suggests some hypothesises about this mechanism.

* Cette expérience a été conduite en 1969 par P. Jospa et M. Wajskop. Les premiers résultats furent traités par R. Deschamps et l'ensemble de ce travail a ensuite été repris par R. Beeckmans.

I. INTRODUCTION.

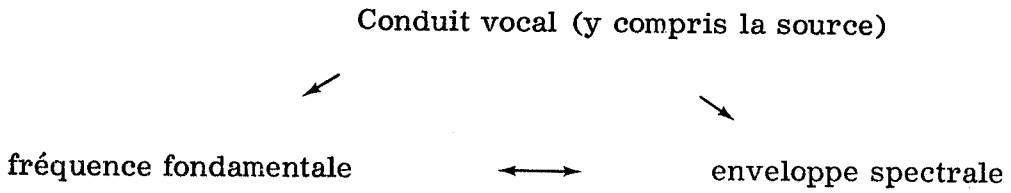
Les différences formantiques entre voyelles émises par des voix d'hommes et de femmes sont attribuées à des facteurs anatomiques (Fant, 1966), tels que la longueur du conduit vocal et le déplacement du larynx . Par ailleurs, la réduction du conduit vocal chez la femme n'est pas répartie de manière uniforme : elle affecte davantage la cavité pharyngale que la cavité buccale. De plus, la fréquence fondamentale d'une voix de femme est en général supérieure d'une octave à celle d'une voix d'homme.

Une étude récente (Carré, 1968) effectuée à l'aide d'un analogue électrique de la cavité vocale a montré qu'il existait une forte analogie entre une voix de femme et une voix d'homme dont la fréquence fondamentale était augmentée, à la fois sur le plan acoustique (déplacements analogues des formants) et sur le plan anatomique (tendance du conduit vocal de l'homme à reproduire celui de la femme). Cette hypothèse se base sur l'accord satisfaisant obtenu en comparant les déplacements formantiques en 3 situations :

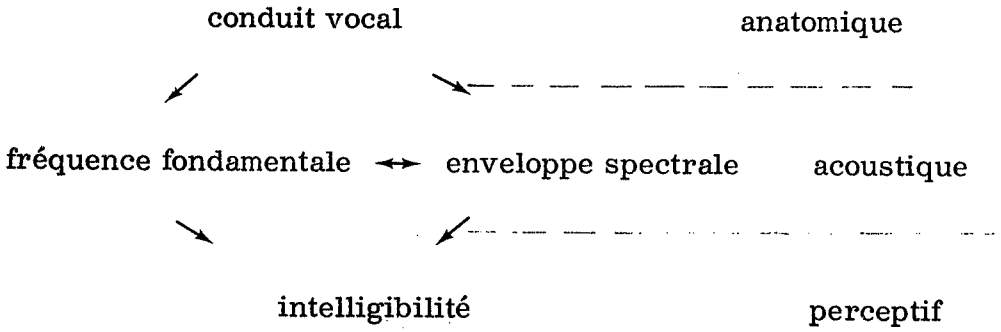
1. voix d'homme → voix de femme
2. voix d'homme à fréquence fondamentale normale (100 Hz) → voix d'homme à fréquence fondamentale élevée (200 Hz).
3. voix synthétique obtenue par simulation du conduit auditif vocal de l'homme → voix synthétique obtenue en réduisant le conduit de 2 cm au niveau du début de la cavité pharyngale côté larynx.

Toutefois la seule réduction du conduit vocal, suffisante pour les voyelles antérieures (ɛ), (e), (i) n'aboutissait qu'à un accord médiocre en ce qui concerne les voyelles postérieures (u), (ɑ), (ɔ), (a). Ceci était dû au fait que la relative rigidité du larynx lors de son élévation, provoquait une modification de la constriction principale. Un ajustement supplémentaire, consistant à déplacer la constriction de 1 cm ou à réduire sa section fournit, pour ces voyelles postérieures un bon accord avec les deux premières situations.

Il semble donc que les changements de l'appareil vocal imposent une relation entre la fréquence fondamentale et l'enveloppe spectrale (cf. schéma 1)



Il était tentant de vérifier si cette relation au niveau acoustique résultant du niveau anatomique pouvait inférer sur le niveau perceptif (cf. schéma 2)



L'occasion nous en a été offerte par le groupe d'analyse-synthèse de Grenoble qui désirait contrôler l'intelligibilité de la production fournie par son appareillage de simulation vocale. Si l'hypothèse émise plus haut est correcte, des voyelles synthétiques de fréquences formantiques, correspondant à celles de voyelles de voix d'homme, devraient voir leur intelligibilité décroître avec l'augmentation de la fréquence fondamentale.

Il était intéressant par ailleurs d'étudier les résultats par une méthode d'échelle multidimensionnelle; en effet, les études récentes s'appuyant sur ces méthodes ont permis de déterminer que deux ou trois dimensions au plus pouvaient rendre compte de l'espace perceptif des voyelles, ces dimensions étant dans certains cas, interprétables en fonction des formants. Parmi l'ensemble de ces méthodes, la méthode INDSCAL offrait l'avantage d'étudier l'évolution de cet espace perceptif, en fonction de l'augmentation de F_0.

2. METHODE EXPERIMENTALE.

-2.1. Stimuli : sept voyelles : i , e , ɛ , a , u , o , et ɔ ont été synthétisées à l'aide de l'analogue électrique de la cavité vocale de l'F.N.S.C. de Grenoble (Carré, 1968). Pour éviter les phénomènes de réduction temporelle, leur durée a été normalisée à 500 ms. avec des fronts d'ouverture et de fermeture exponentiels.
Chacune de ces voyelles a été produite, son enveloppe spectrale étant maintenue constante, aux sept fréquences fondamentales suivantes : 80-100-120-150-200-250 et 350 Hz.
A ces 49 (7x7) stimuli artificiels ainsi obtenus furent ajoutées 14 voyelles naturelles prononcées par une voix d'homme (f_o = 100 Hz) et une voix de femme (f_o = 250). Les deux premiers formants de chaque voyelle synthétique étaient ceux de la voyelle de voix d'homme, correspondante.
L'ensemble des stimuli a été copié en ordre aléatoire sur bande magnétique; l'intervalle séparant chaque voyelle était de 8 secondes. Cinq bandes de 63 (49 + 14) stimuli furent ainsi préparés.

-2.2. Sujets : Huit sujets, âgés en moyenne de 21 ans ont pris part à l'expérience. Munis d'écouteurs, ils entendirent une fois quatre premières bandes choisies de manière et en ordre aléatoires et deux fois une cinquième, la séquence était de 5-1-2-3-4-5. Ils furent soumis ainsi à l'audition de 378 (63x6) stimuli dont 6 pour chaque voyelle et chaque fondamentale.
Les réponses écrites, obligatoires pour chaque stimulus, étaient ventilées sur les 7 voyelles. Aucun retour en arrière n'était possible.
Avant l'expérience, les sujets avaient été soumis à un entraînement leur permettant de maîtriser la transcription des voyelles en système I.P.A.

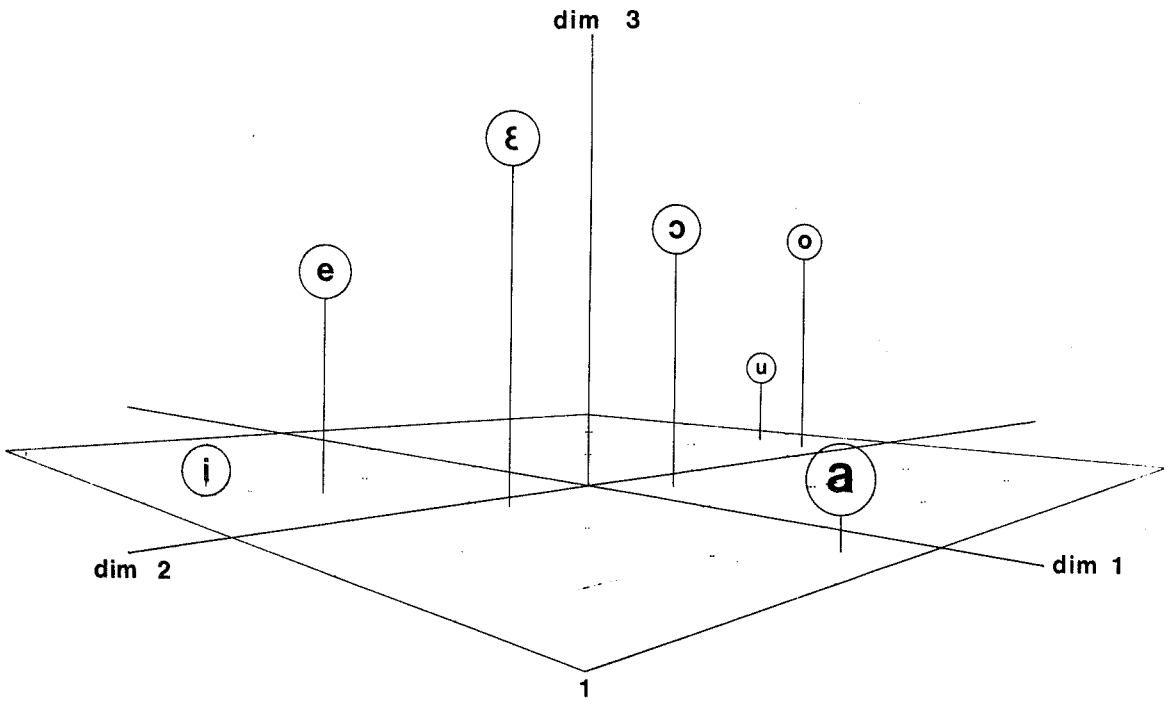
RESULTATS

Le pourcentage global de réponses correctes pour l'ensemble des voyelles décroît avec l'augmentation de F_o. La détermination du coefficient de corrélation basée sur l'information transmise (PEARSON, 1966) vient confirmer ce résultat : différence significative à partir de 200 Hz. Il semble bien que la possibilité de codage diminue lorsqu'on maintient une enveloppe spectrale identique tout en augmentant la fréquence fondamentale.

Ce résultat étant acquis, nous avons appliqué la méthode d'analyse multidimensionnelle INDSCAL (CAROLL, 1972) aux données afin de dégager les dimensions perceptives ainsi que leur évolution en fonction de F_o. Cette méthode présente deux originalités :

1. Elle fournit, à partir de données pour différentes conditions (sujets dans la méthode originale, F_o dans notre cas) un espace général (espace stimuli) dont on peut dériver un espace propre à chaque condition pour l'application de pondérations suivant chacun des axes. L'ensemble de ces pondérations est également représentable (espace conditions).
2. Les axes ainsi déterminés sont univoques et donc, interprétables sans rotation.

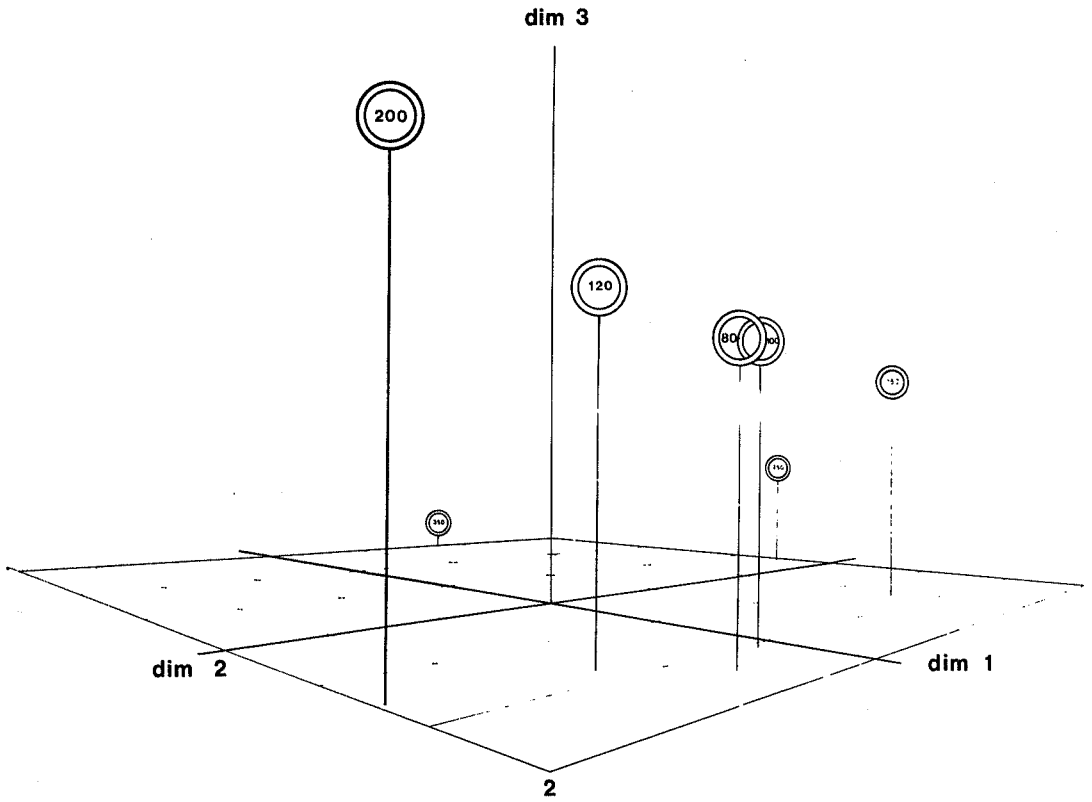
Dans notre expérience, un espace stimuli (graphique 1) à 3 dimensions rend compte de 74 % de la variance totale.

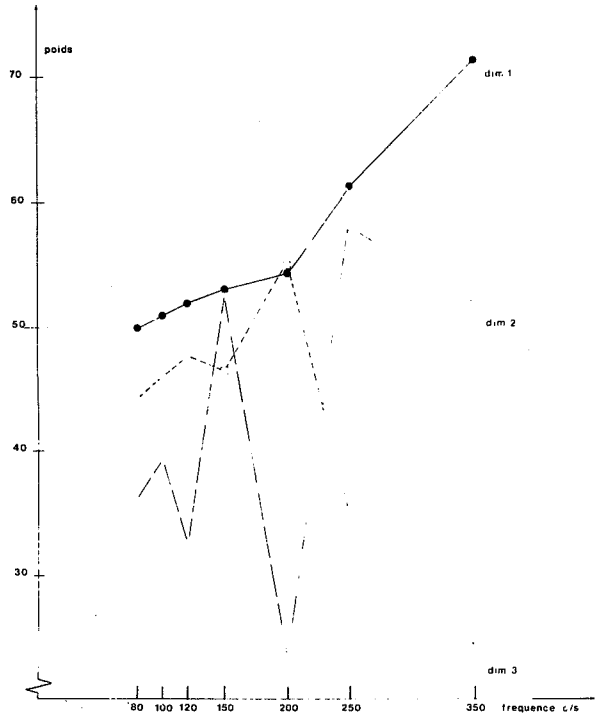


On remarque que les deux premières dimensions sont corrélées avec les positions des deux premiers formants. Afin de tester cette correspondance, nous avons effectué un ajustement entre l'espace physique $F_1 - F_2$ et l'espace perceptif dim1 - dim 2 (graphique 4). On observe une remarquable concordance entre les deux configurations ($r = .99$ pour chaque dimension, signif. à .001); d'autre part les dimensions physiques et psychologiques sont approximativement parallèles.

La troisième dimension n'a pas de relation évidente avec un paramètre physique : elle résulte vraisemblablement de la transformation des données de confusions en données de distances, qui introduit une divergence entre les stimuli extrêmes et moyens.

Les espaces spécifiques à chaque F_o peuvent être dérivés par application à l'espace stimuli des pondérations correspondantes (graphique 2).

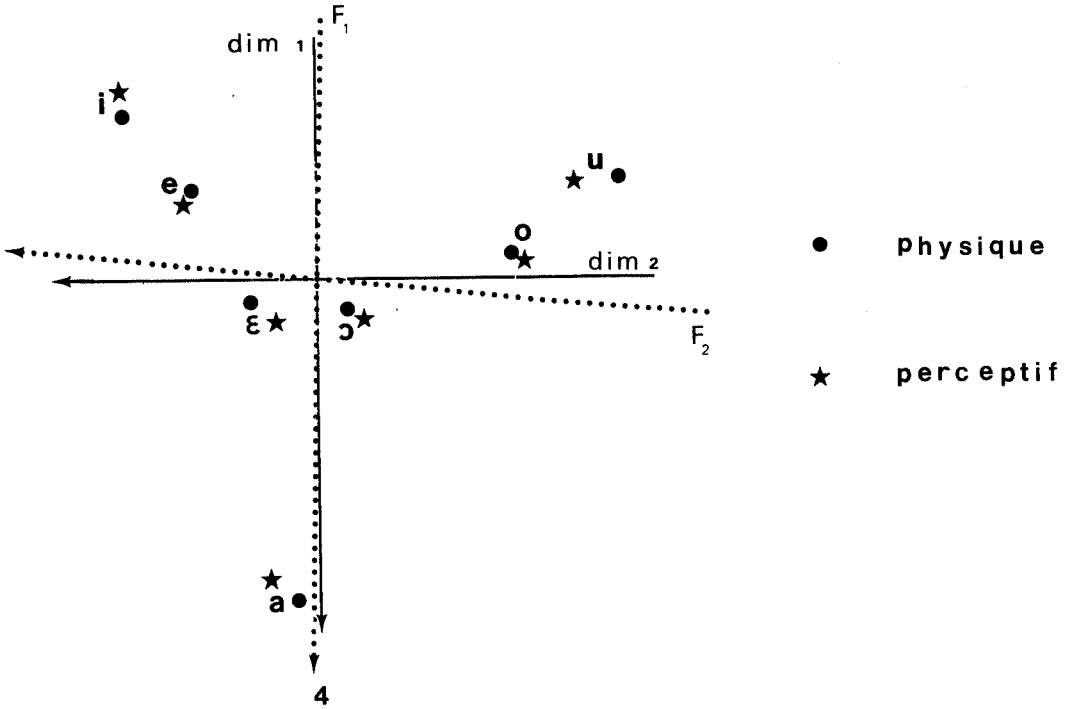




3

Un autre type de présentation (graphique 3), met mieux en évidence l'évolution séparée des poids de chaque dimension en fonction de F_o.

Dans chaque cas, les pondérations sont fonction de la proximité des formants et des harmoniques de F_o. Il apparaît donc que des facteurs liés à la localisation du formant (degré de définition spectrale) déterminent son poids perceptif relatif.



4

CONCLUSIONS.

L'examen des trois résultats principaux
(1) l'augmentation de la fréquence fondamentale entraîne une détérioration de l'intelligibilité
(2) l'application du modèle montre que, au moins, deux critères indépendants sont considérés par les sujets : fréquence des 1er et 2e formants
(3) les sujets sont capables de pondérer chacune des deux informations indépendantes en fonction de leur qualité
suggère quelques hypothèses quant à la manière dont le système perceptif traite l'information formantique en fonction de la fréquence fondamentale. La qualité de l'ajustement entre les 2 axes psychologiques et les 2 axes formantiques impose, par les contraintes mêmes du modèle, que ces deux dimensions soient identiques pour chaque F_o, à un facteur proportionnel près, et en effet, les différences formantiques homme-femmes sont approximativement proportionnelles.

En conclusion, il semble raisonnable de supposer une influence de la fréquence fondamentale à deux niveaux :
(1) influence sur la localisation du formant et indirectement sur l'importance relative de cette donnée sur le plan perceptif.
(2) ajustement (proportionnel) de la dimension formantique.

Les processus décrits ont un effet relativement important ici, mais deux particularités de l'expérience sont à garder en mémoire :
(1) les seules variables physiques caractérisant les 7 voyelles sont F_1 et F_2. Dès lors, qu'elles apparaissent comme dimensions perceptives ne prouvent pas qu'il en serait nécessairement de même si d'autres paramètres physiques devaient intervenir avec leurs fluctuations propres.
(2) La qualité de la localisation des formants et donc leurs poids respectifs revêt vraisemblablement une importance exagérée dans notre cas, à cause de l'absence de micromélodie.

Une expérience actuellement en cours se propose de déterminer à l'aide de stimuli naturels, les limites de ces hypothèses.

Bibliographie.

Carré, R., Lancia, R. & Wajskop, M. Sur la production des voyelles par des locuteurs hommes et femmes. Proceed. 6th Int. Cong. Acoust., 1968, 2, B 31-34

Caroll, J.D. & Chang, J-J. Analysis of individual differences in multidimensional scaling via an N-way generalisation of "Eckart-Young" decomposition. Psychometrika, 1970, 35, 283-319.

Fant, G. A note on vocal tract size factors and non-uniform F-pattern scalings. S.T.L.,Q.P.S.R., 1966, 4, Stockholm.

Hanson, G. Dimensions in speech sound perception : An experimental study of vowel perception. "Ericsson Technics", 1967, 23, 3-175.

Pols, L.C.W., Vander Kamp, L.J.Th & Plomp, R. Perceptual and physical space of vowel sounds. J. Acoust. Soc. Amer., 1969, 46, 458-467.

COMPRESSION ET RECONSTITUTION DE DONNEES SPECTRALES

par M. CARTIER et P. GRAILLOT

C.N.E.T.-LANNION

RESUME : On comprime des données issues d'un vocoder à canaux par analyse factorielle des correspondances.
La réduction de voyelles permet de retrouver des caractéristiques phonétiques. Leur reconstitution demande 3 à 5 paramètres.
Par projection sur des axes adaptés à la parole continue on obtient d'excellents résultats sur les consonnes initiales (D.R.T.). Les voyelles sont relativement plus perturbées (logatomes).

SUMMUARY : Spectral data from a channel vocoder are compressed by means of factorial analysis of correspondences.
Reduction of vowels leads to extraction of phonetic features. For reconstruction of vowels 3 - 5 parameters are needed.
Very good results are obtained on initial consonants (D.R.T.) with axis adapted to continuous speech. Vowels are then relatively more disturbed (non-sense syllabes).

" ... asservir la chair des données
à l'âme des formules... "

J.P. BENZECRI

L'analyse des données

COMPRESSION ET RECONSTITUTION DE DONNEES SPECTRALES

par M. CARTIER et P. GRAILLOT
C.N.E.T. - LANNION

L'analyse factorielle permet de représenter des données multidimensionnelles dans des espaces de dimensions réduites. Elle permet d'extraire mathématiquement un nombre minimum de facteurs indépendants. On peut en déduire une méthode d'extraction automatique de paramètres porteurs d'une information comparable à celle de paramètres tels que les formants. POLS (1) a montré par exemple que l'analyse spectrale de données spectrales de voyelles aboutit à un espace de représentation des voyelles qui peut, moyennant quelques opérations supplémentaires, être corrélé avec l'espace des formants.

La projection des données initiales dans un espace de dimension réduite altère au minimum -au sens de la distance choisie- la qualité des données. On peut donc trouver, à l'intérieur d'une certaine classe de transformations, une formule optimale de compression de données, soit pour les transmettre (KRAMER & MATTHEWS (2), MARANO (3)), soit pour les reconnaître (POLS). L'écoute des données comprimées et reconstituées nous a paru être un bon critère d'évaluation des paramètres obtenus.

Nous présentons des résultats obtenus sur des voyelles isolées et sur un signal de parole réel (essais de netteté), ainsi qu'une discussion en fonction des traits distinctifs.

METHODE EMPLOYEE : L'ANALYSE DES CORRESPONDANCES (4)

Parmi les diverses variantes de l'analyse factorielle, nous avons choisi l'analyse des correspondances qui permet :

- de normaliser en énergie et de réaliser une préaccentuation adaptée au signal traité.
- de projeter sur le même espace les spectres instantanés et les bandes fréquentielles d'analyse, grâce au rôle symétrique joué par les lignes et les colonnes constituant le tableau traité (spectre du signal en fonction du temps).

La distance utilisée est celle du χ^2.

ANALYSE DES DONNEES, AXES DE PROJECTION

Nous avons utilisé les données d'un vocoder à canaux. Cet appareil est en effet parfaitement adapté à la reconstitution de données spectrales :

a. L'analyse de 7 voyelles orales isolées (a, ɛ, i, y, u, o, œ) a donné les résultats indiqués figure 1, où l'on voit les projections des sons sur le plan des deux premiers facteurs. Un son est d'autant plus proche d'un canal que son énergie est concentrée dans ce canal : la projection d'un son est homothétique du centre de gravité des projections des canaux pondérées par les niveaux correspondants.

L'utilisation des coordonnées polaires donne des axes qu'on peut tenter d'interprêter selon un trait grave et un trait de "compacité" (variance). Une présentation légèrement différente, où l'on prend pour coordonnées l'angle polaire dans le plan des deux premiers facteurs et le second facteur (Fig. 2) correspond aux traits grave-aigu et compact-diffus des phonéticiens.

b. L'analyse d'un nombre d'échantillons plus élevé donne des axes relatifs à une parole continue et conduit à la compression de données quelconques.

c. On peut enfin imaginer des présentations plus élaborées des données à comprimer, qui sont actuellement à l'étude. On peut par exemple obtenir un paramètre de segmentation.

RESULTATS

1. La Figure 3 donne les résultats obtenus dans un test à choix fermé sur des voyelles (7 réponses possibles). L'indice des sons présentés aux auditeurs est égal au nombre de facteurs conservés. La voyelle EU nécessite un nombre de facteurs (cinq) plus élevé que les autres voyelles. Pour les six autres, 3 facteurs donnent des résultats aussi bons que les données initiales (5). On pourra comparer aux résultats obtenus par POLS (6) qui a présenté une expérience de reconstitution : le nombre critique de facteurs nécessaires à la reconstitution de voyelles se situe entre 3 et 5.

2. La compression d'un Vocoder à 12 canaux (2200 e.b./s) a été étudiée avec le test de diagnostic par paires minimales (7). On trouvera sur la figure 4 l'évolution du pourcentage de fautes par caractéristique, pour 12 canaux, 5, 3 et 2 facteurs. La dégradation correspondant à chaque réduction est équivalente aux pertes de rapport S/B suivantes:

12 - 5	5 - 3	3 - 2
1 dB	1 dB	2 dB

Rappelons que la perte dûe à la division par deux de la fréquence d'échantillonnage d'un vocoder à canaux est de 2,5 dB (4800 à 2400 e.b./s) (dans le cas d'une écoute limitée à la bande téléphonique) (8). Les consonnes les plus affectées sont les aiguës et les compactes. Le trait de voisement n'est pas touché (5).

3. Enfin une bonne évaluation de la netteté conservée est fournie par un essai aux logatomes réalisé cette fois à partir d'un vocoder à 14 canaux (250 - 4200 Hz) avec un débit initial de 4800 e.b./s.

	14 canaux	5 facteurs	3 facteurs	2 facteurs
% netteté aux logatomes	82	68	52	38
% erreurs voyelles / % erreurs consonnes	0,39	0,51	0,83	0,90

La dégradation observée aux logatomes correspondrait à chaque étape à une une dégradation du rapport S/B de 3 dB (évaluation approximative d'après un calcul d'indice de netteté).

Il est difficile de comparer les trois tests (voyelle, rime, logatomes), dont les conditions étaient différentes. Il est cependant certain que la projection des sons sur des axes adaptés à la parole continue ("axes généraux") détériore relativement plus les voyelles que les consonnes, tandis que la réduction de la fréquence d'échantillonnage n'affecte que les consonnes.

CONCLUSION

Une compression adaptée aux voyelles donne de bons résultats avec 5 facteurs : 3 facteurs suffisent à représenter correctement les 6 voyelles (a, ɛ, i, y, u, o,) après un traitement plus simple qu'une extraction de formants. Les deux premiers facteurs correspondent à une représentation articulatoire (quadrilatère de Straka) et aux traits de gravité et de compacité des voyelles.

Les consonnes sont relativement peu affectées par une projection sur les "axes généraux". Les plus sensibles, par rapport aux caractéristiques initiales du Vocodeur utilisé, sont les consonnes aiguës et les consonnes compactes. Ceci n'est pas contradictoire avec les résultats concernant les voyelles, obtenus avec des axes particuliers. La comparaison des deux essais de netteté montre qu'il est difficile de caractériser certains traitements de la parole par un seul type d'essais.

BIBLIOGRAPHIE

/1/ L.C.W. POLS, L.J.T. VAN DER KAMP, R. PLOMP : "Perceptual and Physical Space of Vowel Souns - J.A.S.A 46 - p. 458-467 (1969).

/2/ H.P. KRAMER, M.V. MATTHEWS : "A linear Coding for Transmitting a Set of Correlated Signals - I.R.E. Trans. I.T. 2 p. 41-46 (1956)

/3/ P. MARANO : "Application de l'Analyse Factorielle des Correspondances à la Compression des Signaux d'Image" : Ann. Télécom. mai-juin 1972 - p. 163-172

/4/ J.P. BENZECRI : "L'Analyse des données " - Dunod (1973).

/5/ P. GRAILLOT : "Compression et reconstitution de données issues du Vocoder" - Analyse et Synthèse de la Parole - Rapport CNET/CRL - Vol. I (1972-1973) - p. 61-70

/6/ L.C.W. POLS : "Analysis and Synthesis of Speech Using a broad-band spectral representation" - Symposium on Speech Perception - Leningrad (Août 1973)

/7/ J.P. PECKELS, M. ROSSI : "Le test de diagnostic par paires minimales" - Revue d'Acoustique - n° 27 - p. 245 - 262 (1973).

/8/ M. CARTIER : "Essais de netteté de niveaux de parole" - Rapport CNET/CRL cité, p. 71-83

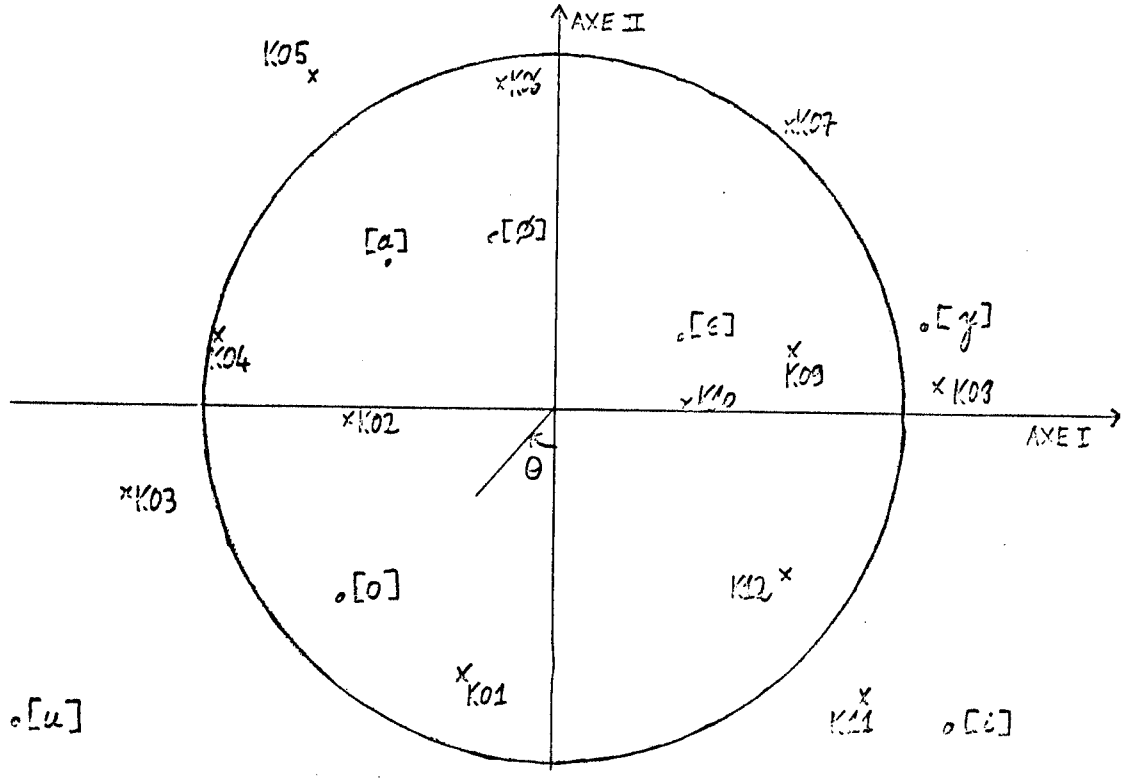


Fig 1 - Projection des 7 voyelles (échantillon central) et des 12 canaux dans le plan des 2 premiers axes factoriels.

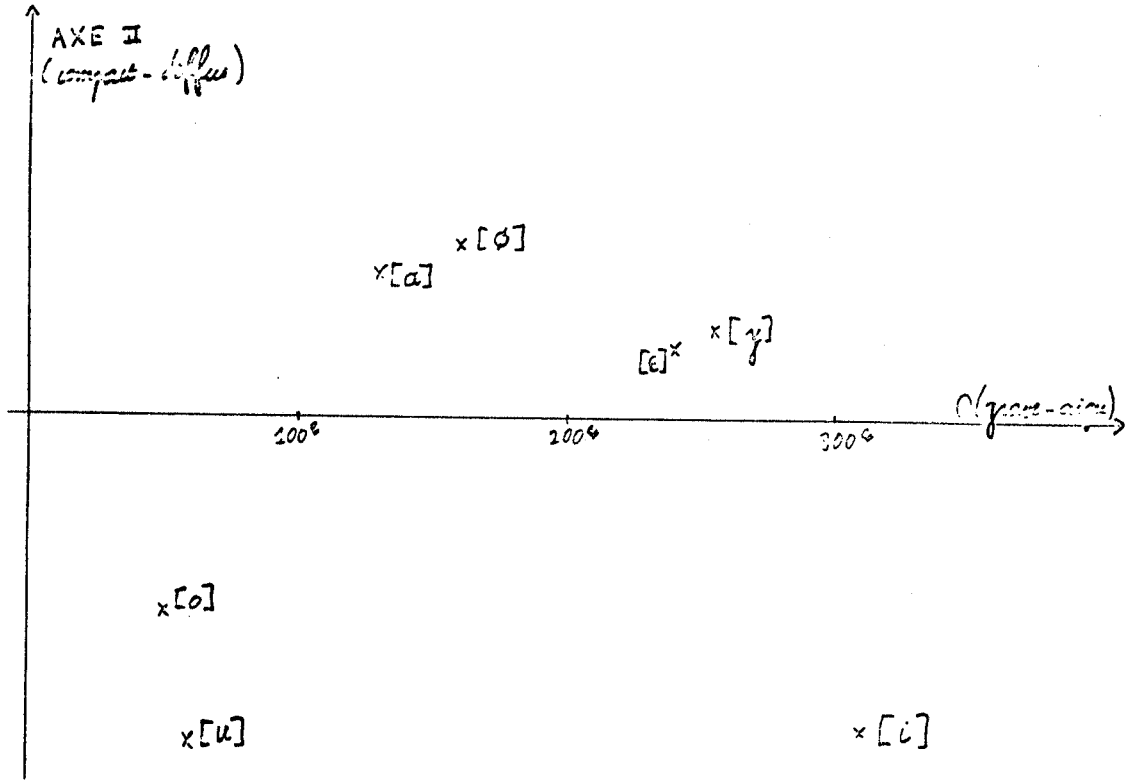


Fig 2 - Projection des 7 voyelles dans le plan formé par le 2ème axe horizontal et l'angle polaire θ.

Nb de confusion / Son prononcé avec	MATRICE DE CONFUSION, chaque son est écouté 144 fois								
	A	O	I	U	OU	È	EU	Nb total de confusions sur 144	% de confusions
A 12	0	0	0	0	0	0	0	0	0 %
A 5	0	0	0	0	0	0	0	0	0 %
A 3	0	5	0	0	0	1	4	10	6,9 %
A 2	0	61	0	0	3	0	0	64	44,4 %
A 1	0	82	0	0	0	1	1	84	58,3 %
O 12	1	0	0	0	6	0	0	7	4,8 %
O 5	0	0	0	0	11	0	0	11	7,6 %
O 3	0	0	0	1	11	0	0	12	8,3 %
O 2	0	0	0	0	5	0	0	5	3,4 %
O 1	8	0	1	0	7	0	1	17	11,8 %
I 12	0	0	0	0	0	0	0	0	0 %
I 5	1	0	0	0	0	1	1	3	2 %
I 3	0	0	0	0	0	0	0	0	0 %
I 2	0	0	0	0	0	5	0	5	3,4 %
I 1	1	1	0	129	0	2	0	133	92,3 %
U 12	0	0	0	0	0	0	1	1	0,6 %
U 5	0	0	0	0	0	1	0	1	0,6 %
U 3	0	0	0	0	0	0	0	0	0 %
U 2	0	0	1	0	0	0	0	1	0,6 %
U 1	0	0	7	0	0	0	1	8	5,5 %
OU 12	2	31	0	0	0	0	0	33	22,9 %
OU 5	0	6	0	0	0	0	0	6	4,1 %
OU 3	0	22	0	0	0	0	0	22	15,2 %
OU 2	8	19	0	0	0	0	2	29	20,1 %
OU 1	9	15	0	1	0	0	1	26	18 %
È 12	0	0	0	0	0	0	0	0	0 %
È 5	0	0	0	0	0	0	5	5	3,4 %
È 3	0	0	0	15	0	0	2	17	11,8 %
È 2	0	1	1	69	0	0	8	79	54,8 %
È 1	0	1	2	51	0	0	4	58	40,2 %
EU 12	0	0	0	0	0	1	0	1	0,6 %
EU 5	0	0	0	0	0	1	0	1	0,6 %
EU 3	35	9	1	0	0	17	0	62	43 %
EU 2	19	4	0	0	2	26	0	51	35,4 %
EU 1	4	101	0	0	9	1	0	115	79,8 %

Fig. 3 Matrice de confusion lors du test des voyelles isolées . Par exemple le U reconstitué avec 1 facteur a été pris, sur 144 écoutes, 7 fois pour un I et une fois pour un EU.

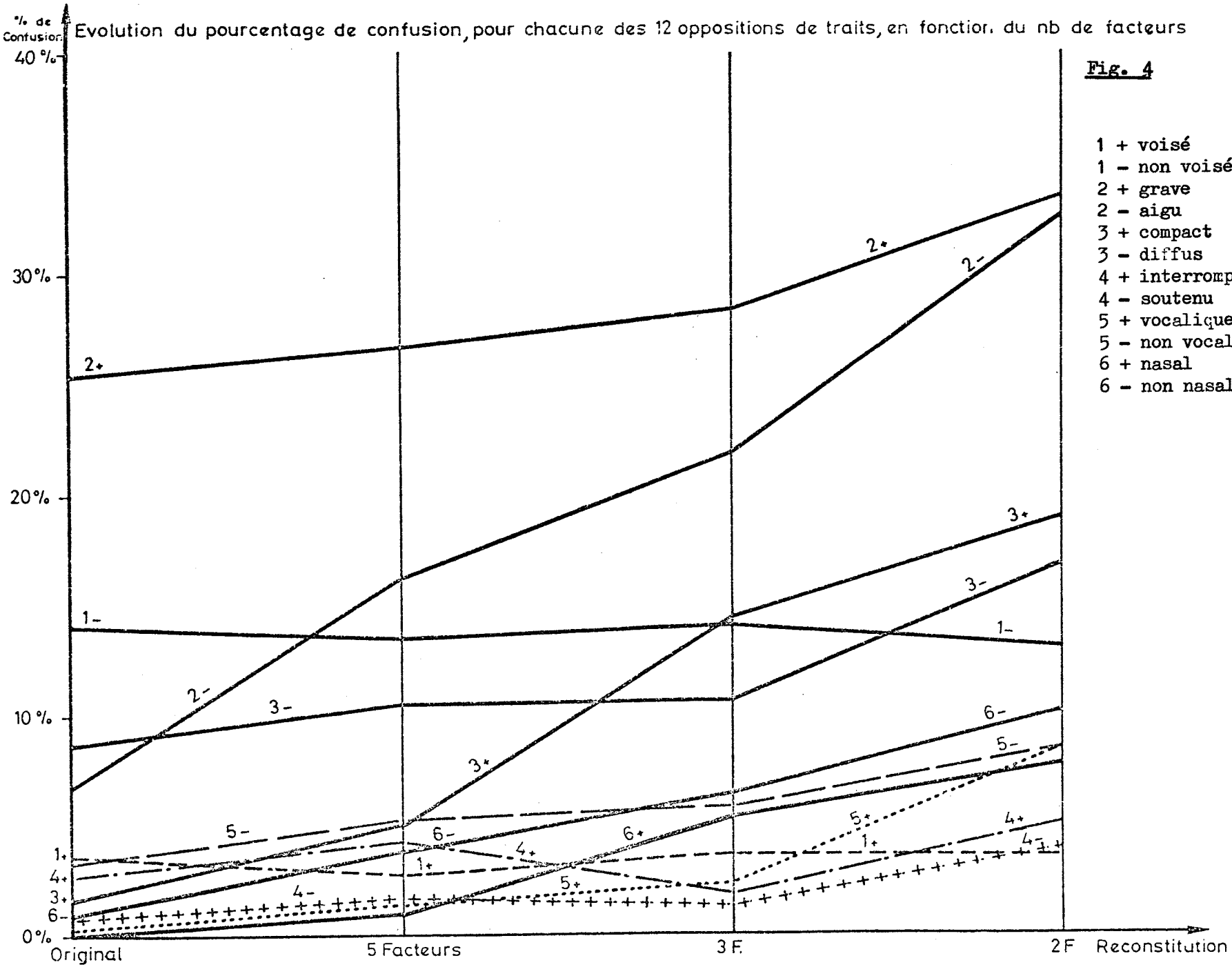

% de Confusion
Evolution du pourcentage de confusion, pour chacune des 12 oppositions de traits, en fonction du nb de facteurs
40 %
30 %
20 %
10 %
0 %
Original
5 Facteurs
3 F.
2 F
Reconstitution
1 + voisé
1 - non voisé
2 + grave
2 - aigu
3 + compact
3 - diffus
4 + interrompu
4 - soutenu
5 + vocalique
5 - non vocalique
6 + nasal
6 - non nasal

Fig. 4

FONCTION D'AIRE DU CONDUIT VOCAL ET ANALYSE ET SYNTHESE DE LA PAROLE

I. EL-MALLAWANY

C.N.E.T.-LANNION

RESUME.

Deux études sont abordées : la détermination de la fonction d'aire du conduit vocal pour des voyelles françaises, et l'analyse et la synthèse de la parole. Les résultats de l'application de quatre méthodes pour la détermination des paramètres d'un filtre numérique linéaire modèle du CV en vue de l'extraction de la fonction d'aire sont illustrés. Ces méthodes sont : le codage prédictif linéaire, les coefficients d'autocorrélation partielle, le filtre de Kalman et le filtre inverse optimal. La deuxième de ces méthodes est incorporée dans un système d'analyse et de synthèse, en raison de la simplicité de l'algorithme, la stabilité du filtre modèle calculé et la facilité d'extraction d'une fonction d'aire qui sert de support de transmission de l'information.

SUMMARY.

Two topics are considered : Vocal tract area function determination for french vowels, and a speech analysis-synthesis system. Four methods have been applied for the determination of the parameters of a digital filter model of the vocal tract and the resulting area functions illustrated. These methods include : Linear predictive coding, partial autocorrelation coefficients, Kalman filter and the Optimal inverse filter. The second of these methods has been incorporated in an analysis-synthesis system because of its simplicity of implementation, the stability of the digital filter models calculated and the ease with which an area function, can be extracted for transmission.

FONCTION D'AIRE DU CONDUIT VOCAL ET ANALYSE ET SYNTHESE DE LA PAROLE

I. EL MALLAWANY

C.N.E.T. - LANNION

La connaissance des défauts [5] propres aux méthodes de compression de la parole à base d'analyse spectrale, a orienté l'intérêt vers l'élaboration d'un modèle dont les paramètres ajustables sont liés à la fonction de transfert du Conduit Vocal (C. V.) et aux caractéristiques de la fonction d'excitation. Une telle étude a trois aspects : la définition de modèles du C. V., d'un modèle de génération de la parole, et de méthodes de détermination des paramètres. L'objet de cette communication est d'aborder deux applications de cette étude, à savoir : la détermination de la fonction d'aire du C. V. pour des voyelles françaises sans nasalités, et la simulation d'un système d'analyse et de synthèse de la parole. Les algorithmes utilisés étant abondamment décrits ailleurs [1-6], ne seront pas détaillés ci-dessous. Néanmoins, nous préciserons dans quelles conditions il convient d'appliquer ces algorithmes et rendrons compte des résultats obtenus.

I. - LE MODELE.

Moyennant certaines hypothèses [4-6] il est possible de rendre compte du fonctionnement du C. V. à l'aide d'une succession de N sections cylindriques d'égale longueur. L'erreur d'approximation étant faible pour des signaux de bande de fréquence ≤ 5 KHz. En tenant compte d'une approximation de la charge aux lèvres et des conditions à la glotte, il est possible de définir une fonction de transfert du C. V. Les configurations du C. V. étant relativement stationnaires (5 à 10 ms), il est possible de définir la fonction de transfert en la variable complexe z.

Les approximations simples de la charge aux lèvres sont au nombre de 3 [6] : le circuit résonateur [8], le circuit parallèle (R et L), la charge résistive. L'ordre étant celui de la meilleure approximation et de plus grande complexité, l'étude du rayonnement des lèvres au capteur établit que la pression au capteur est donnée approximativement par la dérivée de la vitesse volumique aux lèvres [9].

Deux modèles du C. V. [6] sont obtenus en appliquant les conditions de continuité de pression et de vitesse volumique à la jonction entre deux sections. Dans le modèle A la charge adoptée est résistive. Une telle charge n'est pas réaliste mais a le mérite de conduire à un modèle très simple de la forme :

$$H(z) = K/(1 - \sum_{k=i}^{N} a_k \, z^{-k}) \qquad (1)$$

Dans ce cas, les coefficients $\{a_k\}$ sont liés aux aires de sections, A_n, du C. V. par les relations suivantes [4-6] :

$$a_k^{(n-1)} = (a_k^{(n)} - k_n\, a_{n-k}^{(n)}) \,/\, (1 - k_n^2) \qquad (2)$$

et

$$k_n = (A_n - A_{n+1}) \,/\, (A_n + A_{n+1}) \qquad (3)$$

où les coefficients $\{a_k^{(n-1)}\}$ sont les coefficients d'un filtre modèle à $(n-1)^{ème}$ sections obtenu à partir de celui à n sections dont la dernière section a été supprimée et une termination adaptée à la (n-1) ième section.

Le modèle B [6] tient compte d'une charge dite circuit parallèle. La fonction de transfert correspondante est [6] :

$$G(z) = K\,(1 + Qz^{-1}) \,/\, (1 + \sum_{k=1}^{N+1} b_k\, z^{-k}) \qquad (4)$$

$|Q|$ étant < 1, il est possible de ramener G(z) à une fonction ne contenant que des pôles.

Donc un modèle adéquat de représentation du C. V. est un filtre numérique ne contenant que des pôles.

Bien que les études de simulation aient démontré la supériorité du modèle B [6], celui-ci n'est pas applicable au niveau de la détermination de la fonction d'aire. La raison en est que le recours à la transformée bilinéaire s'est révélé nécessaire dans l'élaboration de ce modèle afin d'éviter le recouvrement du spectre [7], et que la distorsion non-linéaire introduite par cette transformée est considérable dans le cas où la fréquence d'échantillonnage est faible (< à 50 KHz). Par conséquent, il n'est fait référence ci-dessous qu'au seul modèle A.

II. - DETERMINATION DE LA FONCTION D'AIRE DU CONDUIT VOCAL.

Tout signal de parole est le produit de trois composantes : le signal de source, la fonction de transfert du C. V. et le rayonnement des lèvres au capteur . Dans le cas des sons sonores la source est située à la glotte et produit des créneaux de périodicité quasi-constante. Le spectre de la source est caractérisé, en moyenne, par une pente décroissante de - 12 dB/octave. Du fait du rayonnement, la pression au capteur est donnée approximativement par la dérivée de la vitesse volumique aux lèvres, soit un spectre de pente croissante par 6 dB/octave. Par conséquent, la détermination du profil d'aire du C. V. suppose l'élimination de ces facteurs du signal avant l'application d'un algorithme menant à la détermination des $\{a_k\}$ dans (1) ; ces $\{a_k\}$ conduisant aux valeurs des aires de section à l'aide de (2) et (3).
Les méthodes utilisées pour le calcul des coefficients du modèle (1) à partir des échantillons de parole dans un intervalle donné sont au nombre de quatre : le Codage Prédictif Linéaire (CPL) [4-6], les Coefficients d'Autocorrélation Partielle (CAP) [1,5,6], le Filtre de Kalman (FK) [2,6], et le Filtre Inverse Optimal (FIO) [3,6].

Il reste à déterminer l'ordre N du filtre modèle. On peut démontrer [4-5] que N doit être tel que son produit avec la période d'échantillonnage T soit égal à deux fois le temps de propagation d'une onde de la glotte aux lèvres. Par conséquent, il est difficile de l'estimer, à priori, car la longueur du C. V., L, varie suivant le locuteur et le phonème prononcé. La connaissance de N est nécessaire à l'obtention d'un profil d'aire correct.

Il reste la question de savoir quelle stratégie adopter en vue de déterminer les $\{a_k\}$ propres à la seule fonction de transfert du C. V. Une approche consiste à adopter un modèle global en y intégrant deux pôles supplémentaires pour tenir compte de la source et du rayonnement. Il en résulte, qu'après la détermination des $\{a_k\}$, il faut calculer les pôles du filtre, opération longue et dont le résultat n'est pas toujours garanti. Il s'ensuit une séparation des 2 pôles de la source qui sont réels ou complexes conjugués, situés proche de l'axe des réels positifs dans le plan Z. Cette approche n'est pas probante.

Une approche différente consiste à situer l'intervalle de fermeture de la glotte dans une période de mélodie, intervalle dans lequel le C. V. est en oscillation libre, et la fonction d'excitation nulle. Néanmoins, il faut intégrer la pression captée au microphone pour obtenir une approximation de la vitesse volumique aux lèvres [9]. Cette intégration est nécessaire, si l'extraction de la fonction d'excitation (créneau glottale) par déconvolution est requise. A défaut, la déconvolution produit un signal à pointes multiples. L'application de cette stratégie en vue de l'extraction du créneau glottal s'est avérée efficace (figure 1), mais les profils d'aire étaient médiocres (figure 2).

Dans le cas où la période de fermeture de la glotte est non nulle, sa localisation dans une période de mélodie n'est pas aisée. Par conséquent, il faut déterminer d'autres conditions d'application qui soient valables dans tous les cas. Une telle approche consiste à prétraiter le signal de parole afin d'y supprimer l'influence de la fonction d'excitation et du rayonnement. Le spectre global de ces deux composantes étant représenté par une pente de - 6 db/octave. Par conséquent, pour annuler en première approximation cette déformation du spectre du C. V., il faut réaliser une préemphase de 6 dB/octave. La question se pose, ensuite, de savoir sur quel intervalle appliquer l'analyse : en synchronisme avec la mélodie ou sur un intervalle constant de 20 ms. Etant donné que les effets de la source et du rayonnement ne sont pas, en fait, éliminés mais substantiellement réduits, le choix se porte sur l'analyse en synchronisme avec la mélodie, car, autrement, la variabilité du modèle peut être relativement importante.

Le moyen le plus simple d'effectuer la préemphase sur le signal de parole consiste à prendre sa dérivée. Néanmoins, la dérivée (pôle au point z = 1 dans le plan Z) ne permet pas d'approcher correctement la pente croissante de 6 dB/octave.

Pour les basses fréquences l'approximation est correcte, mais la pente diminue progressivement pour les hautes fréquences. Cet effet est contraire à celui recherché. En fait, FANT [8] remarque que la fonction d'excitation tend à avoir une pente de plus de - 12 dB/octave pour les fréquences > 3 KHz environ. Par conséquent, cette "égalisation" ne paraît pas suffisante. La solution consiste à "égaliser" adaptativement les effets de la source et le rayonnement à l'aide de trois pôles réels.

Des résultats de l'application de cette approche sont donnés sur les figures 3 et 4, la méthode utilisée étant le filtre inverse optimal. On note que dans le cas de la voyelle "I" [i] les résultats obtenus ne sont pas très sensibles à l'intervalle d'analyse (20 ms ou en synchr. avec mélodie). Par ailleurs, si l'application d'une égalisation adaptative paraît plus adéquate qu'une égalisation fixe, les écarts dans les fonctions d'aires sont faibles. Il en va tout autrement pour la voyelle OU [u]. En effet, l'application d'une égalisation adaptative, conjointement avec un intervalle d'analyse en synchronisme avec la mélodie, conduit à des résultats très largement supérieurs à ceux obtenus dans d'autres conditions. On notera que pour permettre la comparaison des profils d'aire, qui sont obtenus d'une manière relative, avec les données de FANT [8] relatives à des voyelles russes, la valeur maximum de la fonction d'aire est maintenue constante. La fréquence d'échantillonnage est de 10 KHz dans tous les cas.

Il est montré sur la figure 5 un profil d'aire d'un [i], et en hachure la fourchette de variation des valeurs des aires de section pour cette voyelle soutenue ; l'intervalle d'analyse étant de 20 ms. Sur les figures 6 à 10 sont portés les résultats obtenus pour différentes voyelles non-nasales et ce en appliquant les différents algorithmes cités ci-dessus. A ce point, il convient de préciser que dans le cas des méthodes FIO et CAP nous n'avons rencontré aucun problème d'instabilité du filtre modèle (1), c'est une propriété intrinsèque à ces méthodes. Il n'en va pas de même pour les deux autres méthodes : CPI et FK. Il est possible de rendre le filtre stable, soit en calculant les pôles et en les ramenant à l'intérieur du cercle unité, soit en modifiant les valeurs $\{a_k\}$ en $\{a_k\ \alpha^k\}$ jusqu'à obtention de la stabilité ; cette modification a pour effet de ramener les pôles à l'intérieur du cercle unité sans avoir à les calculer.

En général, il n'est pas possible d'obtenir des profils d'aire meilleurs que ceux obtenus dans les mêmes conditions (sans méthode de programmation non-linéaire), en raison du bruit qui entache inévitablement le signal, de la contrainte d'un tube acoustique de longueur quantifiée $\Delta = cT/2$ pour représenter le profil d'aire d'un C. V. dont la longueur est différente, et enfin de la nécessité fondamentale de maintenir la structure formantique du signal avec un faible nombre de paramètres.

III. - ANALYSE ET SYNTHESE DE LA PAROLE.

Dans le cas de sons sonores sans nasalité, la fonction de transfert peut être représentée par un filtre numérique ne contenant que des pôles (1).

Dans le cas contraire (sons nasalisés et sons sourds), il y aura en plus des antirésonances. Ces zéros, situés à l'intérieur du cercle unité, peuvent être remplacés par des pôles (dont le nombre dépendra de la précision requise), ce qui réduit le modèle à un filtre numérique linéaire constitué de pôles exclusivement, de la forme :

$$G(z) = Kg \; / \; (1 - \sum_{k=1}^{N} b_k \, z^{-k}) \qquad (5)$$

La source d'excitation peut être représentée approximativement par deux pôles et le rayonnement des lèvres au capteur par une dérivation. Le tout peut être réduit à un modèle à deux pôles [5]. Il vient le modèle global :

$$H(z) = K_h \; / \; (1 - \sum_{k=1}^{p} a_k \, z^{-k}) \; ; \; p = N+2 \qquad (6)$$

Dans ces conditions, le modèle de génération de la parole prend la forme d'un commutateur qui sélectionne le type de générateur adéquat (suite d'impulsions pour des sons sonores et bruit blanc autrement) pour exciter le filtre numérique (6). Par conséquent, à l'analyse il faut déterminer, sur un intervalle de temps MT, les coefficients $\{a_k\}$. L'approche que nous avons adoptée consiste à appliquer une préemphase fixe au signal de parole, s_n, avant la détermination des a_k. Cette préemphase, qui est approximativement de 6 dB/octave, permet d'égaliser approximativement les effets de la source et du rayonnement, et trouve sa justification en ce que la fonction d'aire qui sera déduite après le calcul des $\{a_k\}$ aura des transitions moins brutales. La notion de fonction d'aire n'a dans ce cas qu'une signification abstraite, elle servira de support pour la transmission de l'information. En fait, la quantification des a_k peut conduire à l'instabilité du filtre à la synthèse, et il vient qu'un codage plus fiable s'applique aux aires, A_n, dont les valeurs ne peuvent être négatives (condition de stabilité du filtre).

Le modèle de génération de la parole retenu est représenté sur la figure 11. Deux générateurs font fonction de source d'excitation, l'un émettant des impulsions d'amplitude et de périodicité variables pour les sons sonores, l'autre un bruit blanc pour les sons sourds. Un amplificateur G permet d'ajuster l'énergie du signal source. Le "conduit vocal" est représenté par un filtre numérique et enfin l'effet global du spectre source plus rayonnement est introduit à l'aide d'un pôle.

L'analyse consiste, par conséquent, à appliquer une préemphase à la séquence s_n sur l'intervalle MT (20 ms), ce qui produit $\{v_n\}$. On applique la fenêtre de Hamming à cette séquence $\{v_n\}$, puis on procède à la détermination des paramètres $\{a_k\}$. L'algorithme adopté est celui dit des Coefficients d'Autocorrélation Partielle [1-5-6] en raison de la garantie de stabilité du filtre numérique calculé (on aurait pu également choisir FIO).

Deux considérations ont prédominé dans ce choix : simplicité des algorithmes et connaissance précise du temps d'exécution de ces algorithmes. Deux conditions essentielles à la réalisation ultérieure d'un équipement spécialisé. La méthode consiste en la détermination des paramètres d'un filtre numérique en échelle modèle du Conduit Vocal (C. V.). Le filtre peut être ramené au modèle (1). Cette méthode permet la détermination très simple des "coefficients de réflection", $\{k_n\}$, entre sections du C. V., d'où la "fonction d'aire" est déduite. Ces aires sont quantifiées et transmises simultanément avec un indicateur de mélodie, la période de mélodie et l'énergie du signal.

La méthode de calcul des k_n décrite ailleurs [1,5-6] suit le cheminement représenté sur la figure 12. Il reste un choix à opérer à savoir l'ordre p du filtre (6). Dans le cas de la parole, il n'existe aucun moyen d'estimer de façon adaptative l'onde p du filtre, en raison du bruit qui entache inévitablement le signal de parole. Néanmoins, la fonction d'aire déterminée et la réponse spectrale du modèle sont très sensibles à l'ordre p. Une trop faible valeur de p entraîne une approximation grossière de la réponse spectrale ; une trop grande valeur conduit à l'approximation de la structure fine du spectre en sus de l'enveloppe spectrale. Par défaut, une valeur moyenne de 17 cm est retenue pour la longueur du C. V, ce qui, compte tenu des considérations faites au § II, conduit au choix de p = 12 pour une fréquence d'échantillonnage de 10 KHz.

La période de mélodie est déduite à partir d'une séquence d'autocorrélation appliquée au signal résiduel issu du filtrage inverse du signal de parole, par l'algorithme "SIFT" [6,10]. La méthode est fiable, et les erreurs de détection peu fréquentes. Certaines erreurs peuvent être corrigées, par exemple, une décision de non-voisement au milieu d'un segment voisé.

La synthèse de la parole suit le schéma fonctionnel de la figure 11, le filtre numérique en échelle peut prendre trois structures différentes mais équivalentes, représentées sur la figure 13. La structure figure 13(c) a été retenue, car elle nécessite un nombre moindre de multiplication. On notera que la structure figure 13(b) est semblable au modèle de Kelly et Lochbaum. Le facteur de gain G est déterminé de sorte que l'énergie du signal synthétisé soit égale à celle du signal à l'analyse dans l'intervalle correspondant. Le calcul tient compte de l'énergie calculée à l'analyse et l'énergie "résiduelle" dans le filtre numérique du fait des intervalles précédents. Les valeurs des coefficients du filtre sont calculées à partir des aires de section suivant l'équation (3).

Dans l'état actuel d'avancement de l'étude, il est possible de dire que ce système d'analyse/synthèse permet de restituer une parole d'une intelligibilité élevée. L'étude de la distribution des aires est en cours en vue de la détermination d'une loi de codage. Cette loi sera logarithmique. En particulier, l'étude portera sur l'opportunité de coder, soit les aires individuellement, soit le rapport des aires de sections contigues, car :

$$k_n = [(A_n/A_{n+1}) - 1] \,/\, [(A_n/A_{n+1}) + 1] \qquad (7)$$

Les conditions d'application du système et les paramètres n'ont pas été optimisés, néanmoins une estimation très réservée du débit en ligne est de 9600 bits/seconde.

Ce système n'a pas les inconvénients des autres méthodes d'analyse/synthèse basées sur le modèle du C. V. Il n'y a ni produit de matrice (p*p), ni la résolution d'un système de p équations linéaires en p inconnues, ni le problème de l'instabilité du filtre modèle, ce qui nécessite un algorithme de détection d'instabilité et un second pour la stabilisation du modèle. Enfin, il n'y a nul besoin d'appliquer la relation récurrente (2) pour parvenir aux coefficients de réflection avant d'en déduire les aires de section.

BIBLIOGRAPHIE

[1] F. ITTAKURA, S. SAITO : "Digital Filtering Techniques for Speech Analysis and Synthesis" - Proc. Int. Congr. Acoust. 7th Budapest (August 1971).

[2] C.J. GUEGUEN, G. CARAYANNIS : "Analyse de la Parole par Filtrage Optimal de Kalman" - Automatisme - Tome XVIII n° 3 - Mars 1973 (pp. 99-105).

[3] H. WAKITA : "Direct Estimation of the Vocal Tract Shape by Inverse Filtering of Acoustic Speech Waveforms" - IEEE Trans. on Audio and Elect., Vol. AU-21, n° 5 October 1973 (pp. 417-427).

[4] B.S. ATAL and S.L. HANAUER : "Speech Analysis and Synthesis by Linear Prediction of the Speech Wave" J.A.S.A. - Vol. 50, n° 2 - Part (2) - 1971 (pp. 637-655).

[5] I. EL MALLAWANY : "Détermination de la fonction d'aire du conduit vocal par codage prédictif". Journées d'Etudes sur la Parole, Groupement des Acousticiens de Langue Française - Lannion (Juin 1972) (pp. 281-306).

[6] I. EL MALLAWANY : "Fonction de transfert et fonction d'aire du C. V." - N. T. CEI/CSI/43 - Rapport d'activité 1973 (1er Février 1974) C. N. E. T. - LANNION.

[7] I. EL MALLAWANY : "Le Filtrage Numérique" - Annales des Télécom. T. 24, n° 3-4 - Mars-Avril 1969 (pp. 89-112).

[8] G. FANT : "Acoustic Theory of Speech Production", Mouton 1970, The Hague.

[9] J.L. FLANAGAN : "Speech Analysis, Synthesis and Perception", Springer-Verlag 1972 New-York - 2nd edition.

[10] J.D. MARKEL : "The Sift algorithm for fundamental frequency estimation" - IEEE Trans. on Audio, Vol. AU-20, n° 5 December 1972.

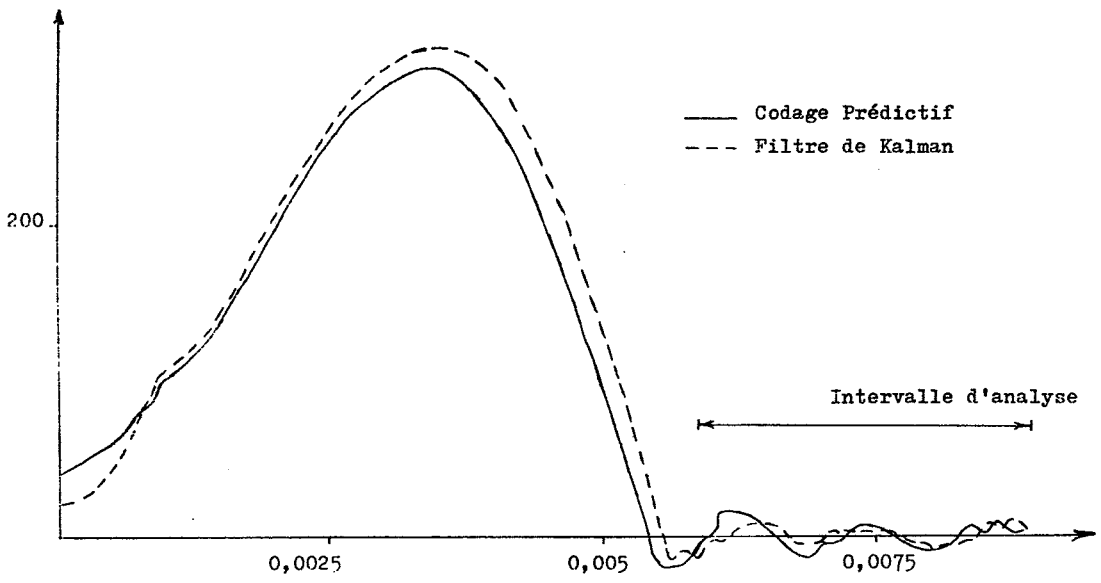


Figure 1 - Déconvolution : créneau Glottal

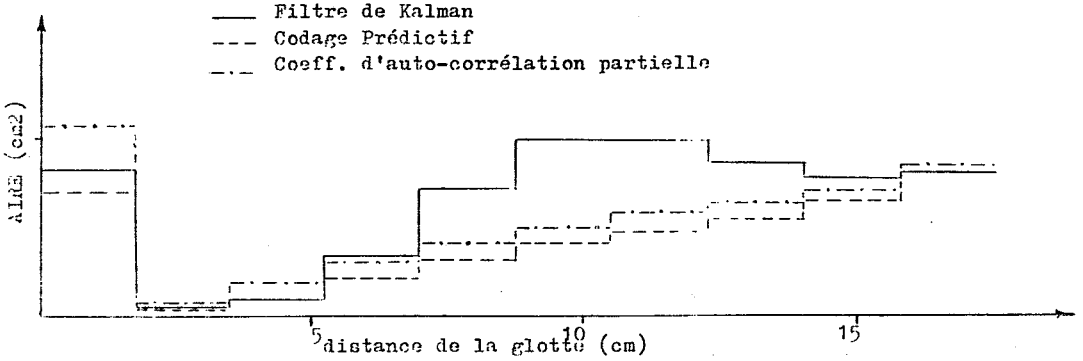
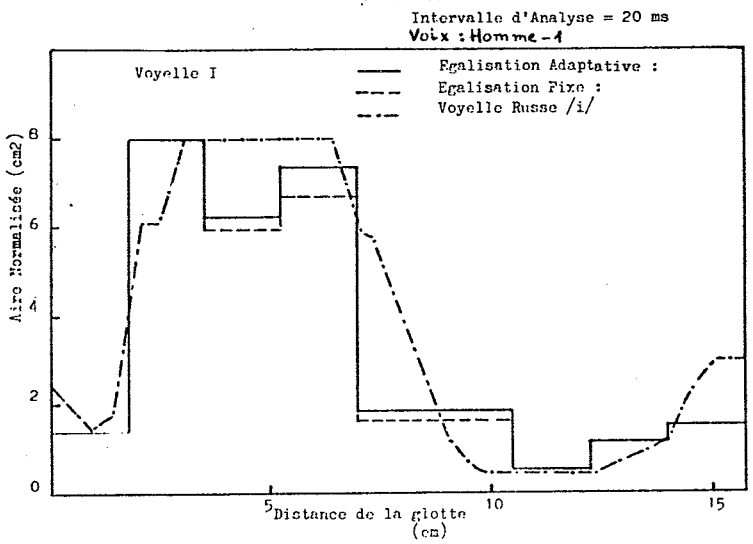


Figure 2 - "O" de Synthèse : Résultats avec T : 10^{-4} secs

Figure 3 : Méthode : Filtre Inverse Optimal



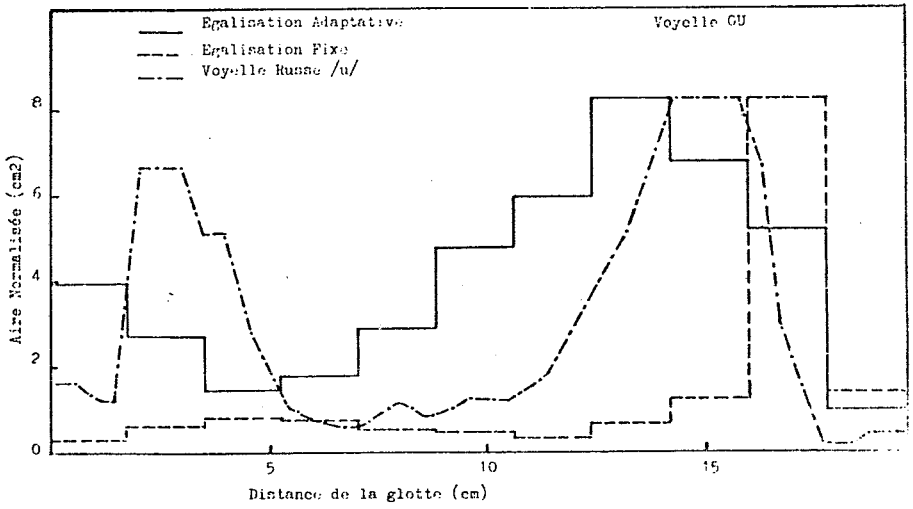
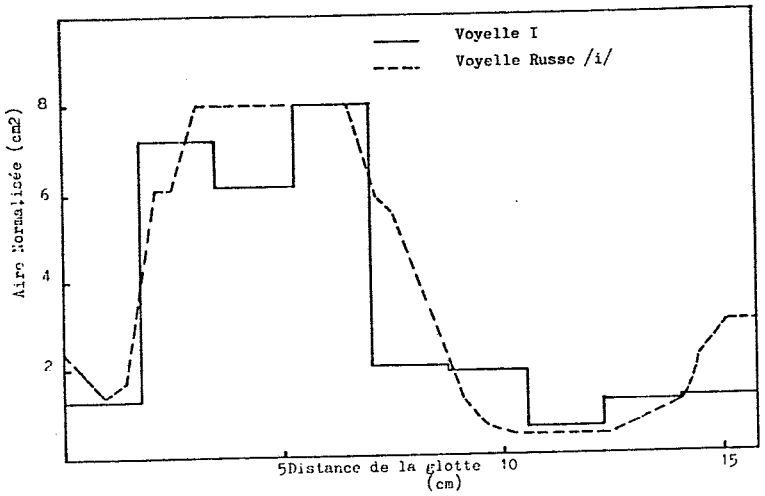
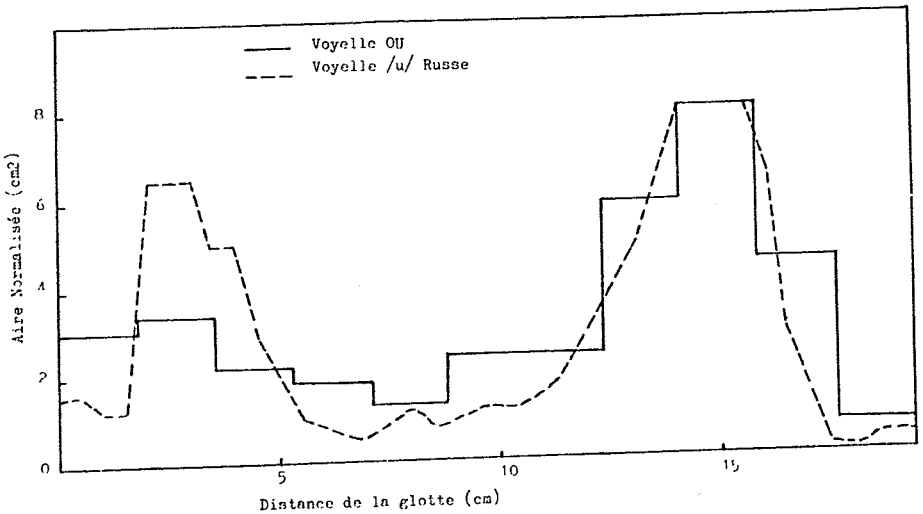


Figure : 4 - Méthode : Filtre Inverse Optimal

Intervalle d'Analyse = période de mélodie

Voix : Homme-1





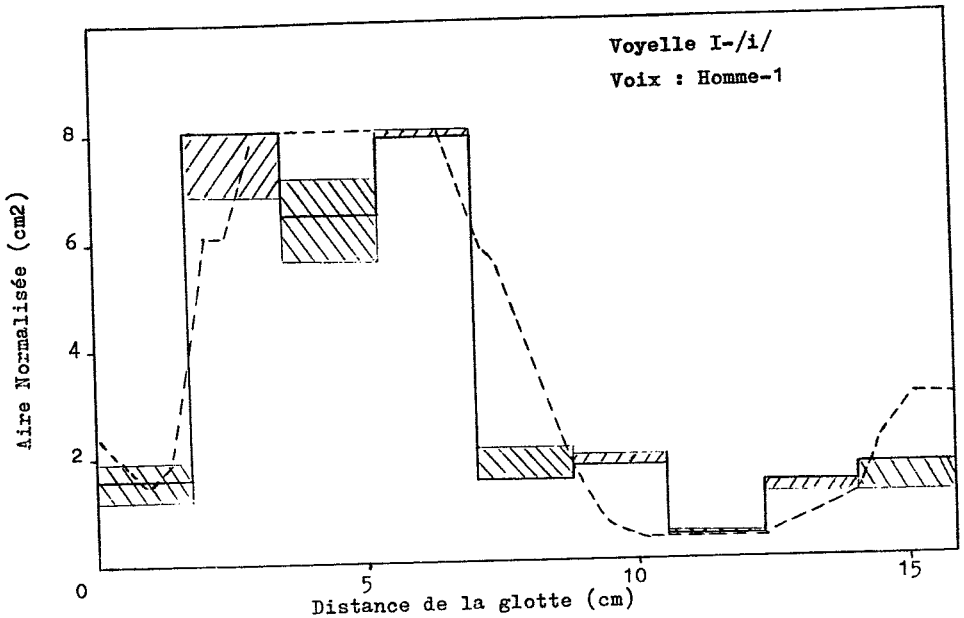


Figure 5 - Méthode : FIO

Intervalle d'Analyse = 20 ms
Egalisation Adaptative

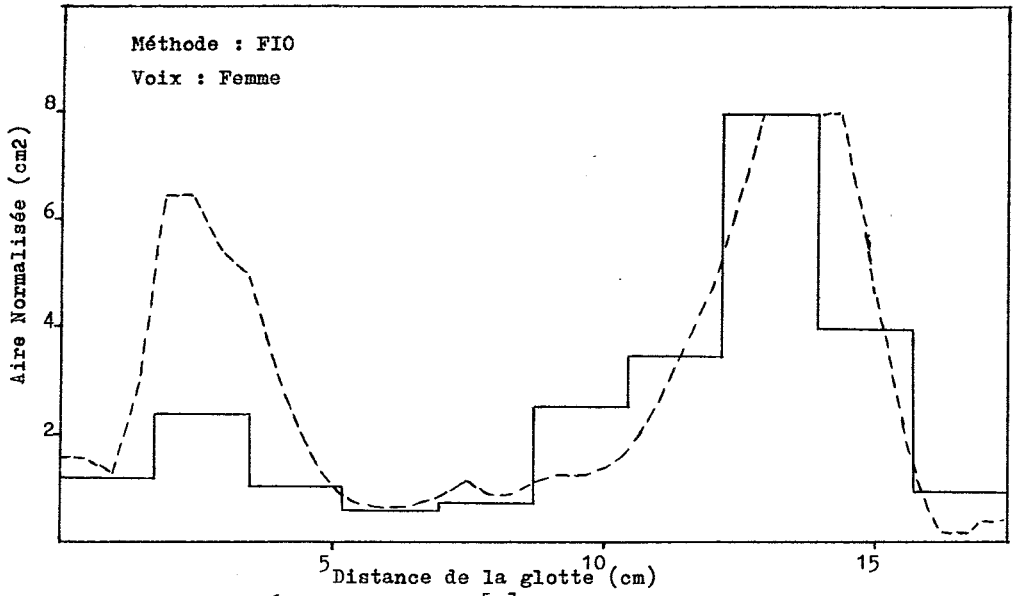


Figure 6 - Voyelle OU - [u]
Intervalle d'analyse = 20 nsec
Egalisation Adaptative
Voyelle [u] Russe

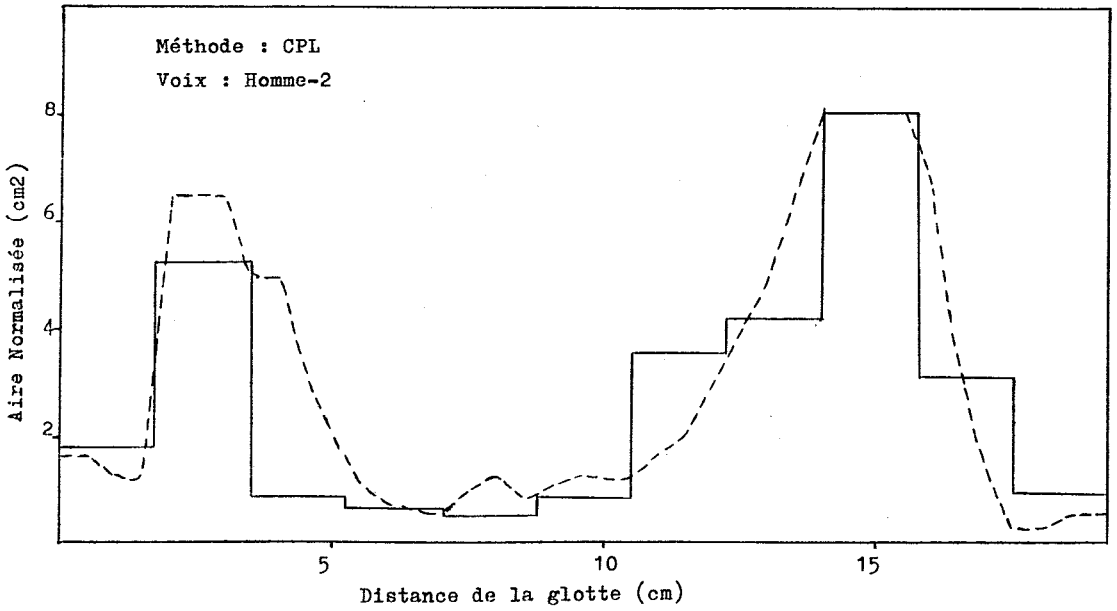


Figure 7 - Voyelle I, [i]
Intervalle d'Analyse : période de mélodie
Egalisation Adaptative
Voix : Homme-2

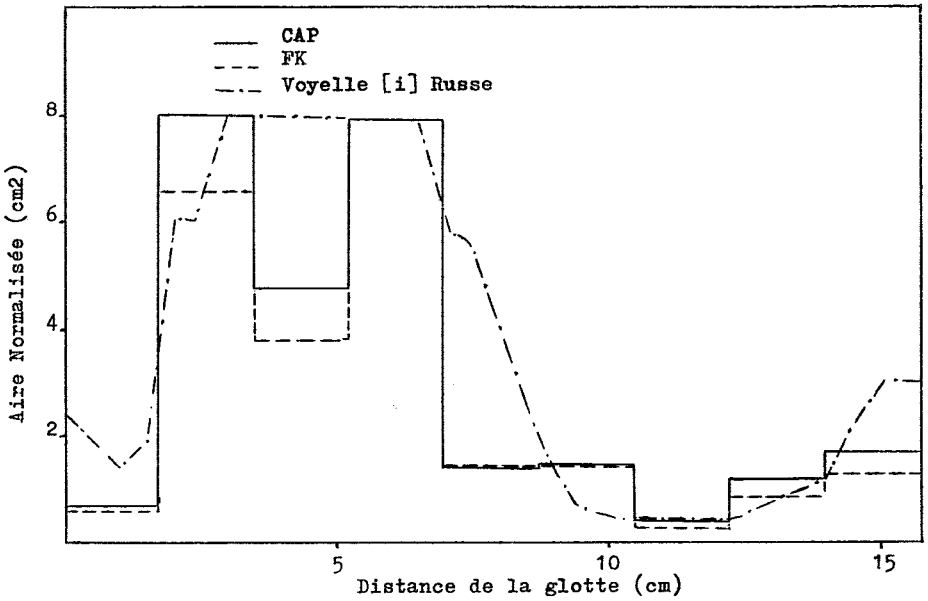


Figure 8 - Voyelle A, [a]
Intervalle d'Analyse : période de mélodie
Egalisation Adaptative
Voix : Homme-2

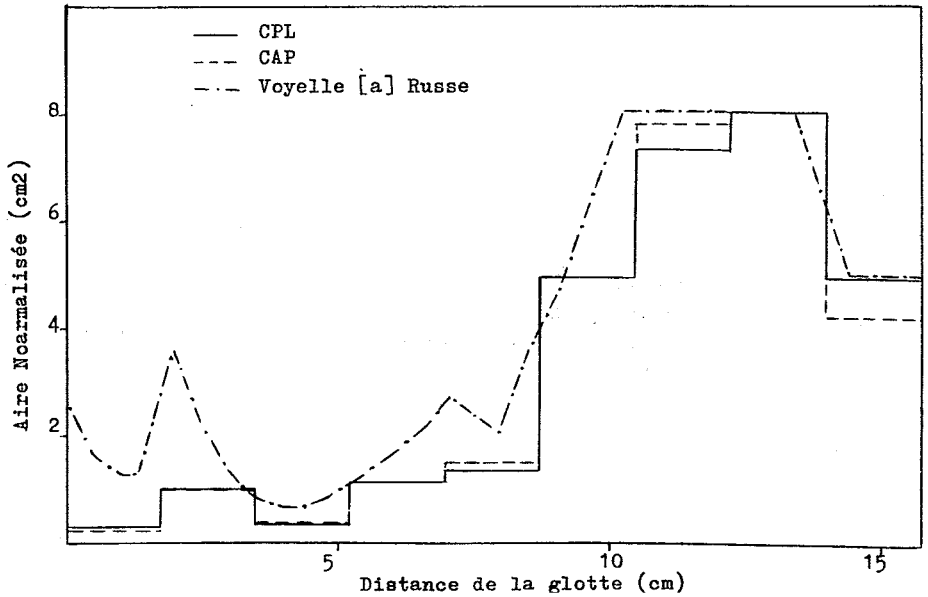


Figure 9 - Intervalle d'Analyse : période de mélodie

Egalisation Adaptative

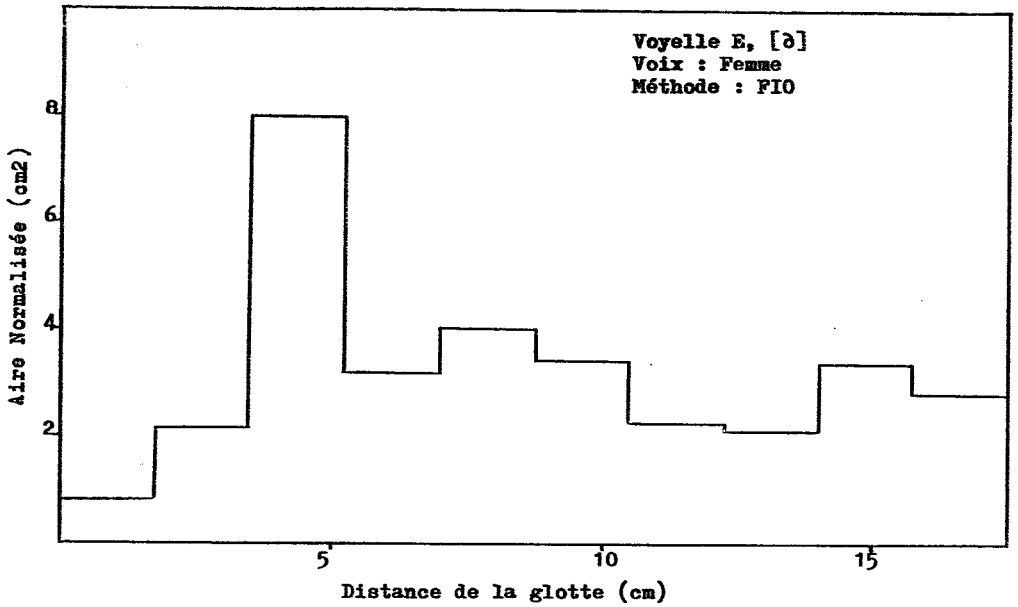
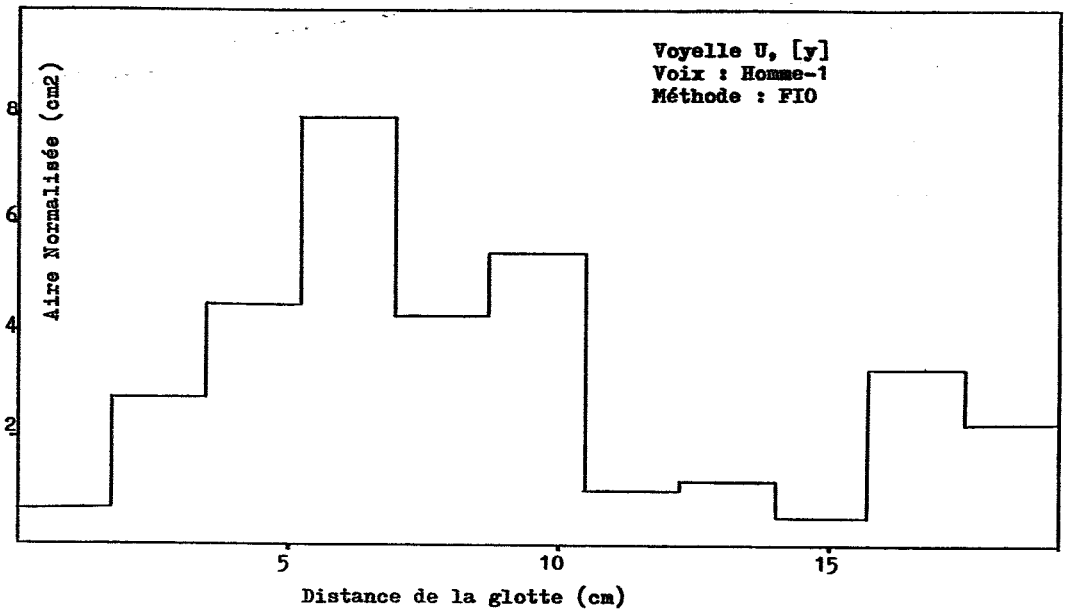
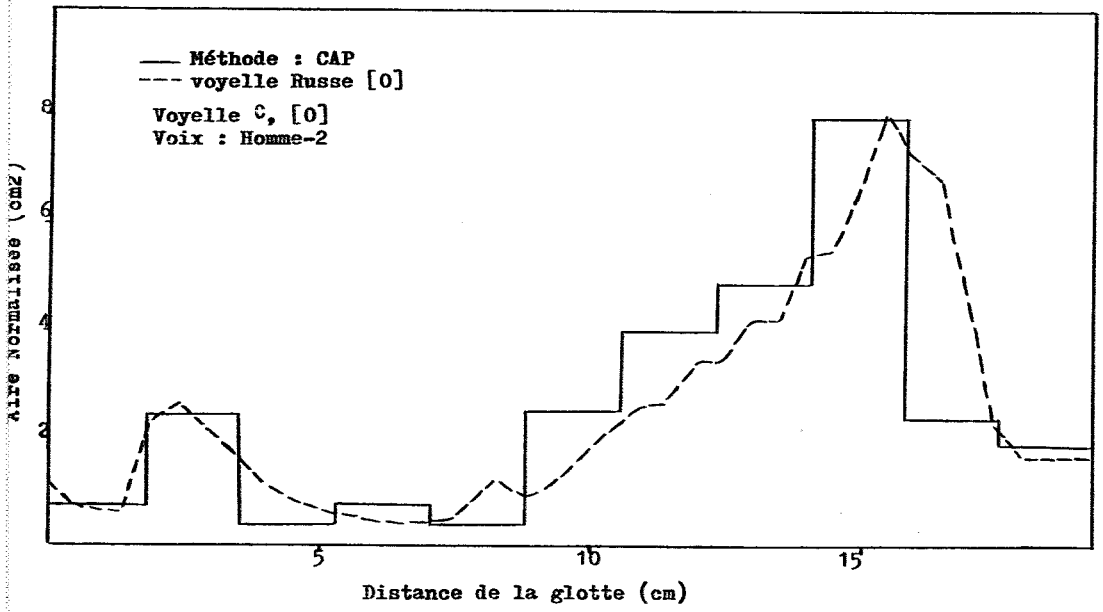
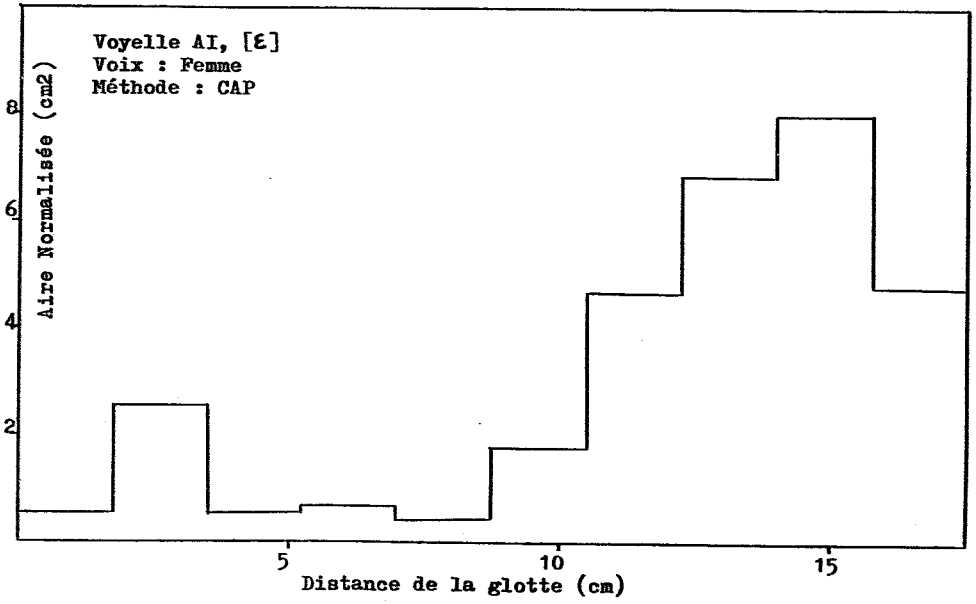


Figure 10 - Intervalle d'Analyse : période de mélodie

Egalisation Adaptative







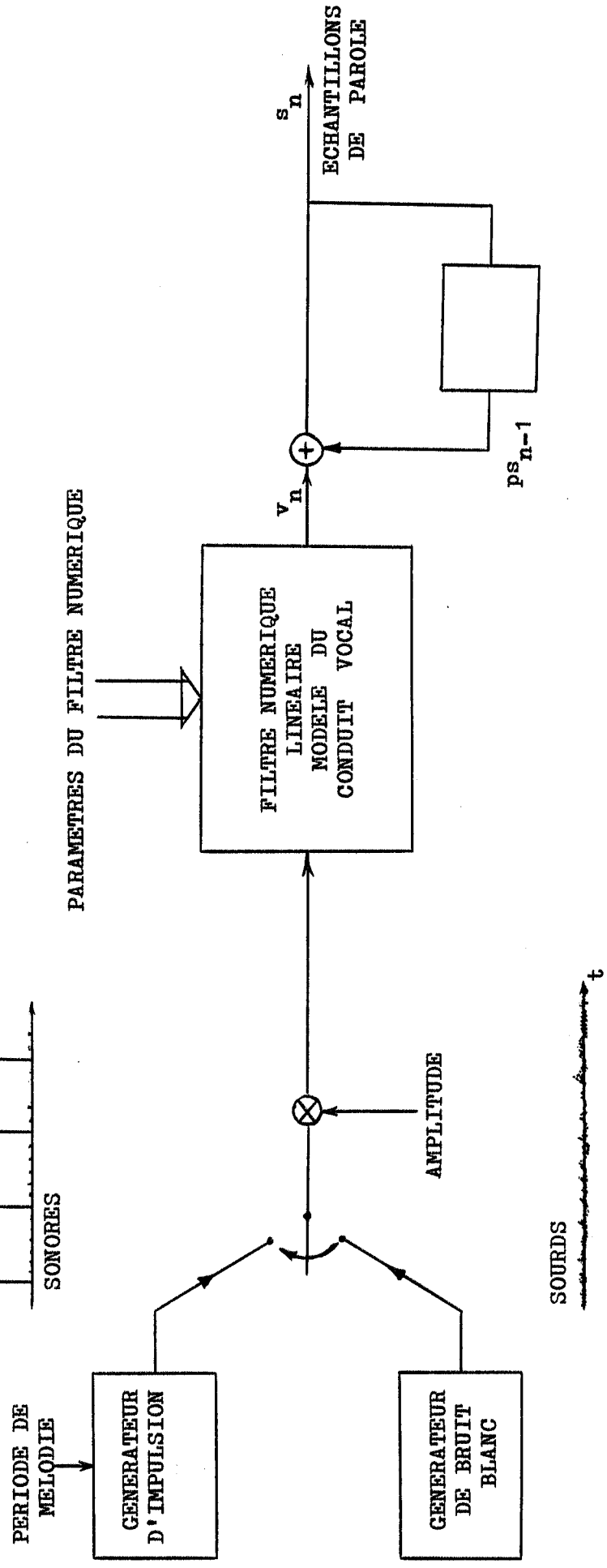


Figure 11 - Modèle numérique de production de la parole

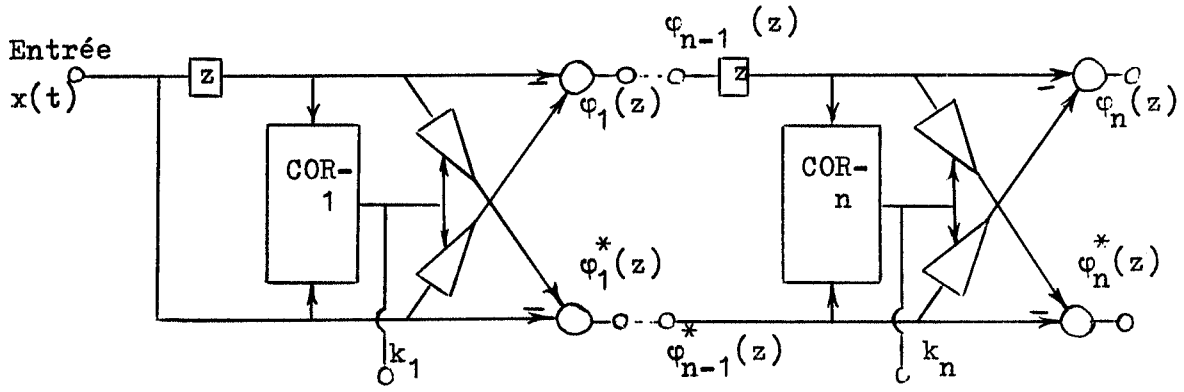
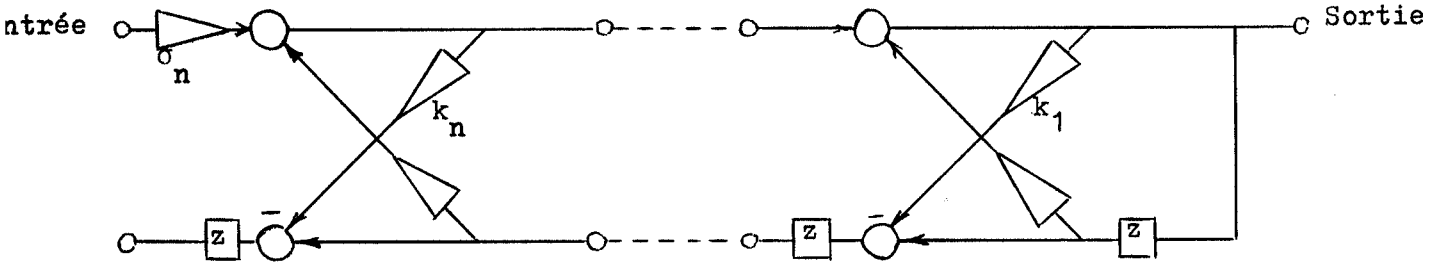
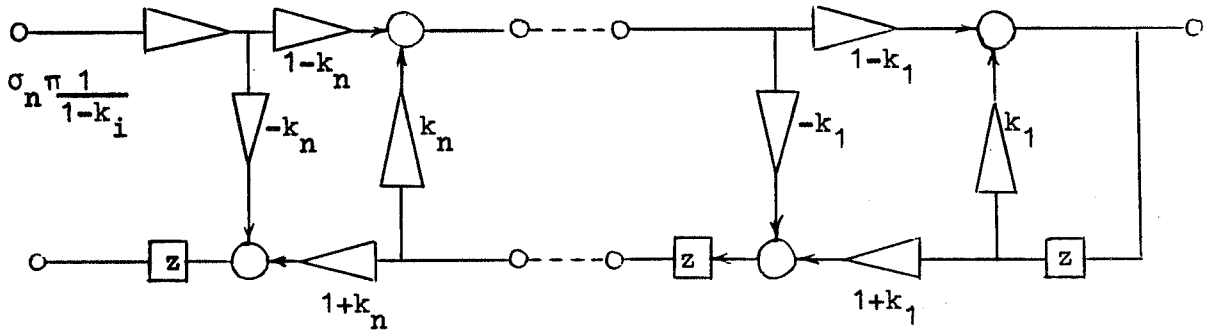


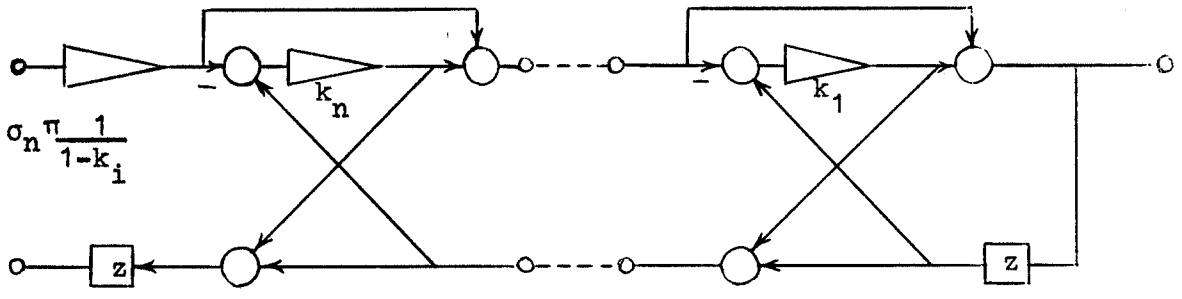
Figure 12 - Filtre numérique adapté pour le calcul des paramètres $\{k_i\}$



a) Filtre numérique à structure en échelle



b) Modèle équivalent [semblable à celui de Kelly]



c) Autre modèle équivalent avec moins de multiplications

Figure 13 - Filtres numériques à structure en échelle pour la synthèse [les filtres sont équivalents]

EFFICACITE DU CODAGE ACOUSTIQUE

Ch.BERGER-VACHON et G.MESNARD
Physique Electronique
Université de LYON I

On envisage les problèmes posés par le codage d'une onde, qu'on peut supposer être acoustique, et l'évolution de l'erreur en fonction du nombre de niveaux.

On peut voir que l'amélioration apportée par le codage évolue comme une progression géométrique.

Les résultats seront ensuite comparés avec ceux donnés par une étude de la divergence entre phonèmes, et avec l'efficacité des voies acoustiques.

EFFICIENCY IN CODING ; SPEECH APPLICATIONS

In this paper, some theorical problems involved in coding a wave are studied (this can be made for an acoustic wave). The evolution of precision is shown, versus the number of levels.

It can be seen the improvement brought by a big number of levels matches a geometric serie laws.

The theorical relationships are then compared with the results given by an experiment on the divergence between phonemes, and with the efficiency of neural pathways of hearing in the brain.

EFFICACITE DU CODAGE ACOUSTIQUE

I - LE CODAGE D'UNE ONDE

1-1 Les probabilités a priori

Soit un alphabet $S = (s_1, \ldots, s_q)$ composé de q symboles. On définit par $I(s_i) = \log_2 \frac{1}{P(s_i)}$ en bits (1) l'information apportée par la réalisation du symbole s_i (où $P(s_i)$ = proba a priori du symbole s_i)

En effet, si le symbole s_i est certain, on n'est pas "plus informé" en apprenant que s_i s'est réalisé.

L'information amenée par S est en moyenne :

$$H(S) = \sum_{i=1}^{q} P(s_i) \log_2 \frac{1}{P(s_i)} \qquad (2) ;$$

H(S) définit l'entropie de la somme S.

Si on recherche les valeurs de $P(s_i)$ qui maximisent H(S), compte-tenu de la contrainte :

$$\sum_{i=1}^{n} P(s_i) = 1 \qquad (3)$$

on peut montrer [1] que H(S) est maximal pour les valeurs de s_i telles que :

$$P(s_i) = \frac{1}{q} \quad \forall\, i \qquad (4)$$

Donc, si nous considérons une onde et que nous disposons de n niveaux pour la coder (par exemple avec un convertisseur analogique digital), les signaux issus du codage transporteront l'information maximale lorsque nous réaliserons la condition (4). Si on connait la répartition de l'énergie dans une bande de fréquence, l'équation (4) permet de délimiter de façon optimale les plages qui seront représentées par un même signe à la sortie du vocodeur, (figure 1).

1-2 Les niveaux de moindre erreur

Pour chacune des plages, on se propose de calculer la valeur qui la représente au mieux (le niveau). L'écart entre le signal x et son niveau v_i après quantification est :

$$\varepsilon = x - v_i \qquad (5) \qquad \text{si } x \in \text{la } i^e \text{ plage.}$$

La valeur moyenne $\overline{\varepsilon^2}$ du carré de l'erreur est, pour la i^e plage, donnée par :

$$\overline{\varepsilon^2} = \int_{u_{i-1}}^{u_i} (x - v_i)^2 \, f(x) \, dx \qquad (6)$$

u_i et u_{i-1} sont les bornes de la i^e plage.

Si on optimise (6) par rapport à v_i on obtient [2], [3] ;

$$\frac{d\overline{\delta^2}}{d v_i} = 0 \quad \Longrightarrow \quad v_i = \frac{\int_{u_{i-1}}^{u_i} x \; f(x)\, dx}{\int_{u_{i-1}}^{u_i} f(x)\, dx} \qquad (7)$$

v_i est donc l'espérance mathématique de la probabilité dans la i^e plage. En exprimant v_i à l'aide de l'équation (7), $\overline{\delta^2}$ s'écrit :

$$\overline{\delta^2} = \sigma^2_x - \sigma^2_v \qquad (8)$$

où

$$\sigma^2_x = \int_{-\infty}^{+\infty} (x - \overline{x})^2 \; f(x)\, dx \qquad (8 \text{ bis})$$

σ^2_x est la variance du signal.

$$\sigma^2_v = \frac{1}{q} \sum_{i=1}^{q} (v_i - \overline{v})^2 \qquad (9)$$

σ^2_v est la variance liée aux niveaux de codage. De plus, on sait que $\overline{x} = \overline{v}$ (th.de la moyenne statistique).

σ^2_x étant constant, plus σ^2_v sera grand, meilleur le codage sera. L'équation (8) implique aussi que $\sigma^2_v \leq \sigma^2_x$.

1-3 Influence du nombre de niveaux sur le codage

Etudions la quantité σ^2_v dans le cas particulier suivant (fig.2). On connait la densité de probabilité de l'énergie à l'aide d'un histogramme. Les limites ont été placées de telle façon que $S_1 = S_2 = S_3 = S_4 = S_5 = S_6$, (cette configuration correspond à 6 intervalles de codage). La densité de probabilité étant uniforme dans chacun des intervalles, le point v_i représentant le i^e intervalle est au milieu de la base du i^e rectangle.

Si on améliore cette décomposition en doublant le nombre des intervalles de codage et en respectant l'équation 4, le point v_i devient une limite de bande ; le i^e rectangle se scinde en deux. Les deux nouveaux rectangles issus du i^e intervalle sont codés par les niveaux

$$v_{1i} = v_i - \frac{\Delta L_i}{4} \quad \text{et} \quad v_{2i} = v_i + \frac{\Delta L_i}{4} \qquad (10)$$

L'équation (9) devient :

$$(\sigma^2_v)_{2q} = \frac{1}{2q} \sum_{i=1}^{q} \left((v_{1i} - \overline{v})^2 + (v_{2i} - \overline{v})^2 \right)$$

$(\sigma^2_v)_{2q}$ représente le terme σ^2_v s'il y a 2 q niveaux, ces 2 q niveaux étant formés à l'aide des q premiers niveaux.

$$(\sigma^2_v)_{2q} = \frac{1}{q} \sum_{i=1}^{q} \left[(v_i - \overline{v})^2 + \frac{\Delta L_i^2}{16} \right] \qquad (11)$$

$$(\sigma^2_v)_{2q} - (\sigma^2_v)_q = \frac{1}{q} \sum_{i=1}^{q} \frac{\Delta L_i^2}{16} = \frac{1}{16} \left(\overline{\Delta L_i^2} \right) \qquad (12)$$

On voit que l'amélioration amenée par le doublement des niveaux de codage est proportionnelle à ΔL_i^2 .

Donc, si on part d'une répartition à q niveaux, une multiplication par deux du nombre des intervalles de codage améliore la précision selon la formule 12. Comme σ^2_v est borné supérieurement par σ^2_x on peut tracer la courbe ; figure 5.

1-4 Forme asymptotique de l'amélioration

L'amélioration obtenue lorsque le nombre de niveaux passe de q_o à 2 q_o (où q_o est un codage de base) est donnée par :

$\Delta_1 = (\sigma^2_v)_{2q_o} - (\sigma^2_v)_{q_o}$; de nouvelles divisions des intervalles initiaux donnent :

$\Delta_2 = (\sigma^2_v)_{4q_o} - (\sigma^2_v)_{2q_o}$ soit :

$$\Delta_k = (\sigma^2_v)_{2^k q_o} - (\sigma^2_v)_{2^{k-1} q_o} \qquad (13)$$

Peut-on étudier Δ_k ?

$$16\,\Delta_k = \frac{1}{q} \sum_{i=1}^{q} \Delta L_i^2 \qquad (14)$$

avec $q = 2^k q_o$

Si on continue à découper les intervalles (fig.2) on améliore de Δ_{k+1} la précision en passant de $2^k q_o$ à $2^{k+1} q_o$ niveaux.

$$16\,\Delta_{k+1} = \frac{1}{2q} \sum_{i=1}^{q} \left[\left(\frac{\Delta L_i}{2} \right)^2 + \left(\frac{\Delta L_i}{2} \right)^2 \right]$$

$$= \frac{1}{4q} \sum_{i=1}^{q} (\Delta L_i)^2 = \frac{16\,\Delta_k}{4}$$

soit $$\Delta_{k+1} = \frac{\Delta_k}{4} \qquad (15)$$

La somme des améliorations évolue donc comme la somme d'une série géométrique, de premier terme $(u_s - u_i)^2$ et de raison $\frac{1}{4}$. Elle converge selon la forme indiquée figure 5.

On peut aussi étudier Δ_k pour q fixé. On sait que :

$$\frac{1}{q} \sum_{i=1}^{q} \Delta L_i^2 = \frac{1}{q} \sum_{i=1}^{q} (\Delta L_i - \overline{\Delta L})^2 + \frac{\left(\sum_{i=1}^{q} \Delta L_i \right)^2}{q^2}$$

soit $16\,\Delta_k = \mathrm{Var}\,\Delta L + (\overline{\Delta L})^2$ (16)

où $\overline{\Delta L}$ est la moyenne des quantités ΔL_i (i varie de 1 à 9) ;

$\overline{\Delta L} = \frac{u_s - u_i}{q}$ est une constante pour q fixé.

Les quantités $(\Delta L_i - \overline{\Delta L})^2$ seront minimales lorsque les valeurs ΔL_i seront égales $\forall i$.

Cette condition sera compatible avec l'équation 4 lorsque la densité de probabilité aura la forme :

$$f(x) = 0 \quad \text{si } x < a$$

$$f(x) = \frac{1}{\rho} \quad \text{si } a \leqslant x \leqslant a + \rho$$

$$f(x) = 0 \quad \text{si } x > a + \rho$$

où a et ρ sont 2 paramètres donnés (figure 4).

II - APPLICATION A LA PAROLE

2-1 Compromis à effectuer au niveau du codage

Considérons un vocodeur à canaux ; on peut en schématiser son fonctionnement figure 5 : l'onde incidente est analysée par un ensemble de filtres délimitant des bandes de fréquences. L'énergie contenue dans chacune des bandes est analysée toutes les τ (en général τ = 20) millisecondes par un convertisseur analogique-digital (C.A.D.).

Supposons que chacun des C.A.D. dispose de q niveaux. Si on connait la répartition de l'énergie f(x), pour un ensemble de phonèmes, dans chacun des canaux, on peut régler les intervalles des C.A.D. selon l'équation 4. Les sorties S_1, S_2, ..., S_r garderont alors le maximum de leur information d'entrée ; si on utilise toutes les possibilités de codage dans tous les canaux, on perd le rapport de l'énergie. Comme on sait (équation 15) qu'à partir d'un certain stade, la multiplication des intervalles de codage n'amène plus qu'une information infime, il est peut-être préférable de ne pas utiliser toutes les possibilités des CAD pour garder l'information sur l'énergie.

Essayons de préciser cette notion :

Nous avons étudié l'efficacité des canaux du vocodeur du CNET à Lannion [4] pour effectuer des séparations de phonèmes selon des critères gaussiens ; figure 6 [5].

Le vocodeur du CNET conserve l'information sur l'énergie totale : les possibilités du codage sont donc plus étendues sur les canaux basse fréquence qui codent entre 0 et 15 ; les canaux haute fréquence codent entre 0 et 6.

On admettra, que pour l'ensemble des 25 phonèmes utilisés, la densité de l'énergie f(x) est à peu près constante quel que soit le niveau.

Peut-on au niveau des séparations, utiliser les formules démontrées précédemment ?

On remarque, par exemple, que le taux de séparation lié aux canaux haute fréquence, fait apparaître une diminution sensible de l'efficacité ; en effet, ce taux varie entre 0,70 et 0,75 si on compare les résultats des canaux 1 et 12 avec des seuils différents (1,80 et 1,98) pour la séparation ; mais il faut compléter cette expérience pour s'assurer que de tels résultats ne sont pas dus seulement aux propriétés spectrales de la parole. Si ce n'est pas le cas, l'allure générale des séparations observées sur ces canaux haute et basse fréquence confirme la forme de l'équation 15.

2-2 Efficacité de l'audition

On considère [5] les propriétés anatomo-physiologiques de l'organe de Corti ; cet organe est constitué par environ 30000 cellules ciliées (25 000 sont situées à l'extérieur du tunnel de Corti et 5 000 sont du côté de la columelle). A ces 30 000 cellules ciliées sont attachés à peu près le même nombre de neurones. La connexion entre les cellules ciliées et les neurones est très compliquée (un neurone est "attaché" à plusieurs cellules ciliées et une cellule ciliée est en connexion avec plusieurs neurones) et sa signification fonctionnelle est encore mal connue. Néanmoins, on sait que les fibres nerveuses ont des significations fonctionnelles différentes ; celles qui passent en périphérie du ruban de Reil latéral transportent des informations sur les sons aigus tandis que celles qui passent à l'intérieur sont relatives à des fréquences plus basses ; la membrane basilaire effectue une première discrimination.

Si on considère qu'un neurone transmet en tout ou rien des impulsions [6] sur une fréquence de récurence de 100 Hz (à cause de la période réfractaire) l'organe de Corti peut transmettre environ 3 000 000 de bits par seconde soit 3 M bauds. Ces possibilités sont nettement supérieures au débit de la parole qui a été évalué à 50 Kbauds environ. Le codage d'une onde est indiqué figure 7.

Quelle est la part efficace dans cette information ? Un auditeur distrait pourra ne retenir pratiquement rien. Sur un plan strictement informationnel une impulsion sur 60 suffit pour transmettre la parole. Compte-tenu des possibilités d'adaptation du cerveau et compte-tenu du fait que les derniers niveaux n'améliorent pratiquement pas l'information, les formules que nous avons établies montrent donc que l'information utile transmise par les voies acoustiques est certainement plus de 60 fois redondante. De telles remarques peuvent être utiles en pathologie.

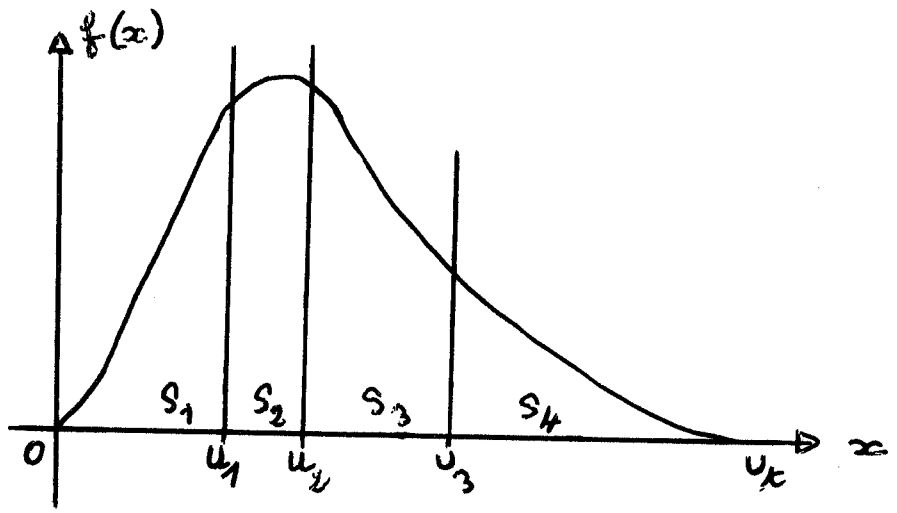


figure 1 : Soit x l'énergie dans un canal et f(x) sa répartition. Les limites U_1, U_2 et U_3 des bandes énergétiques qui seront codées par un seul niveau sont données par l'égalité $S_1 = S_2 = S_3 = S_4$ (cas de 4 niveaux).

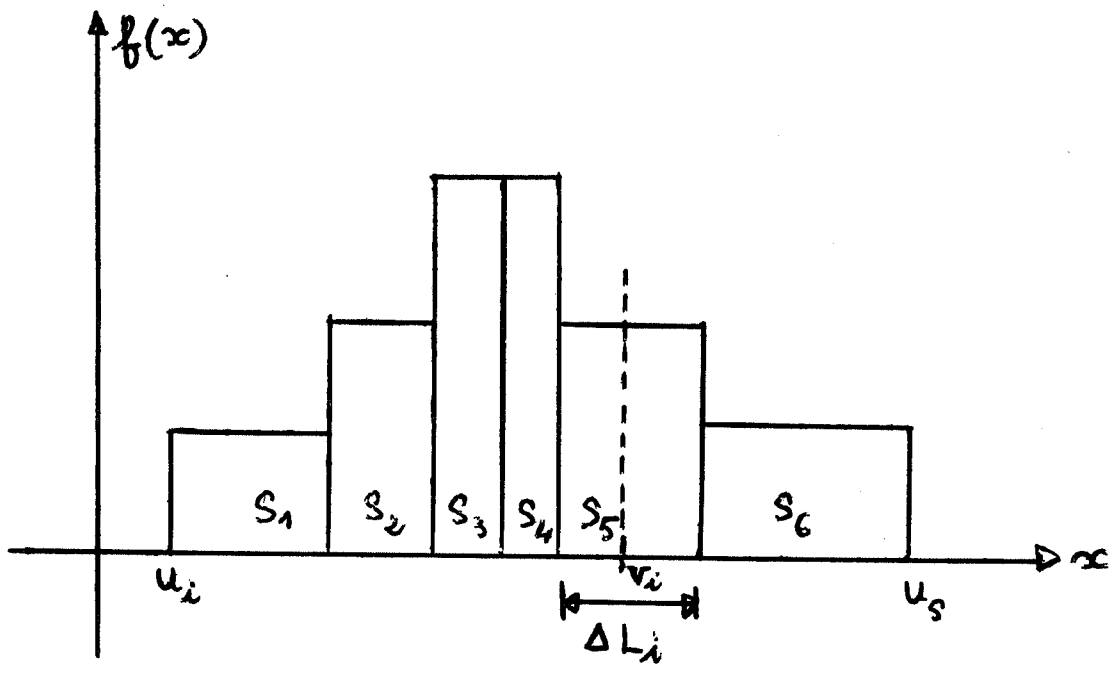


figure 2 : Approximation d'une densité de probabilité à l'aide de rectangles de surface égale.

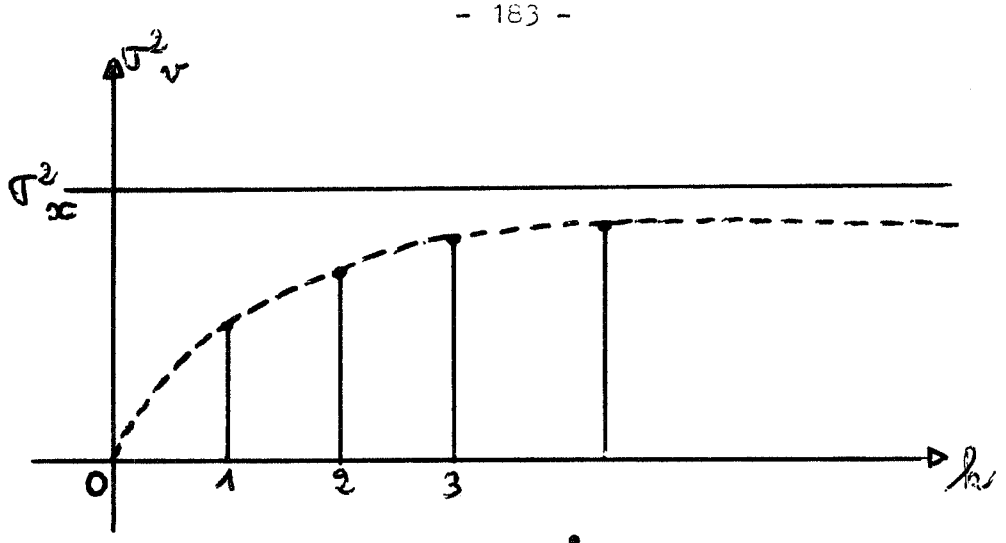


figure 3 : Forme asymptotique de σ^2_v, lorsque le nombre de niveaux devient 2^k.

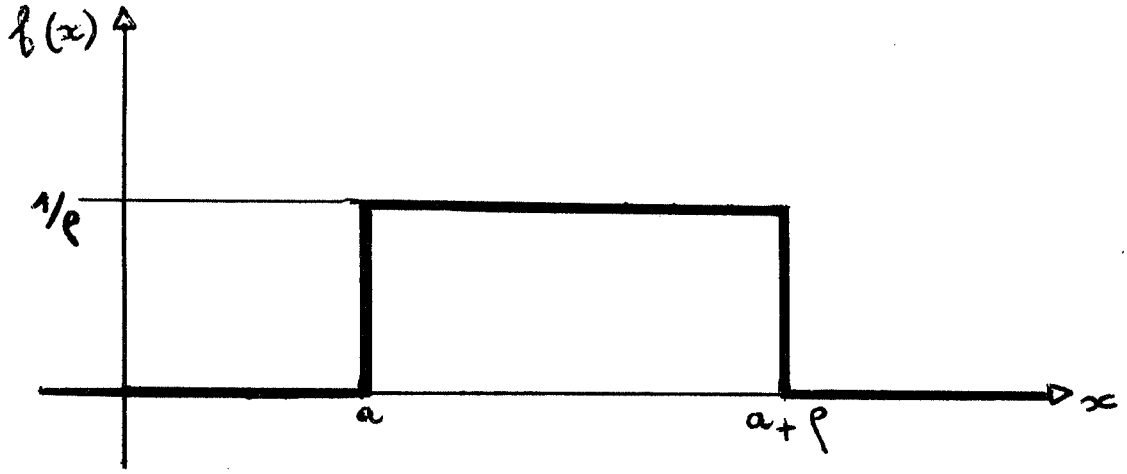


figure 4 : Densité de probabilité uniforme dans l'intervalle [a, a + ρ]

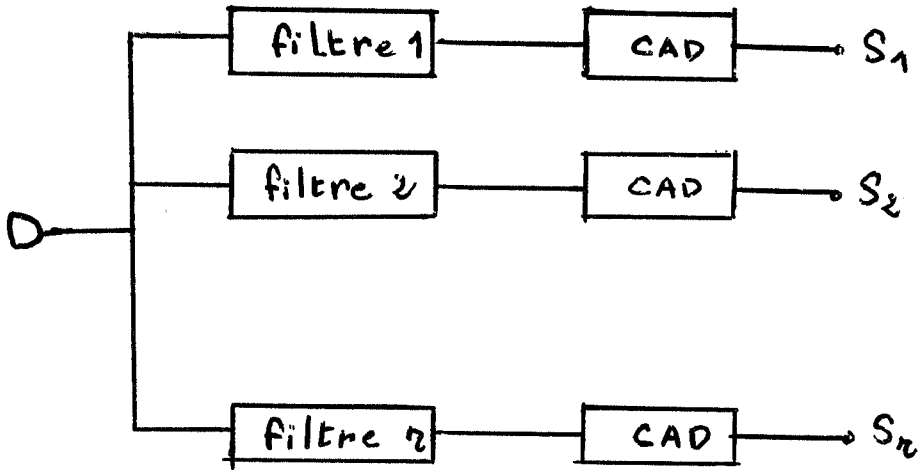


figure 5 : Schéma d'un vocodeur à canaux.

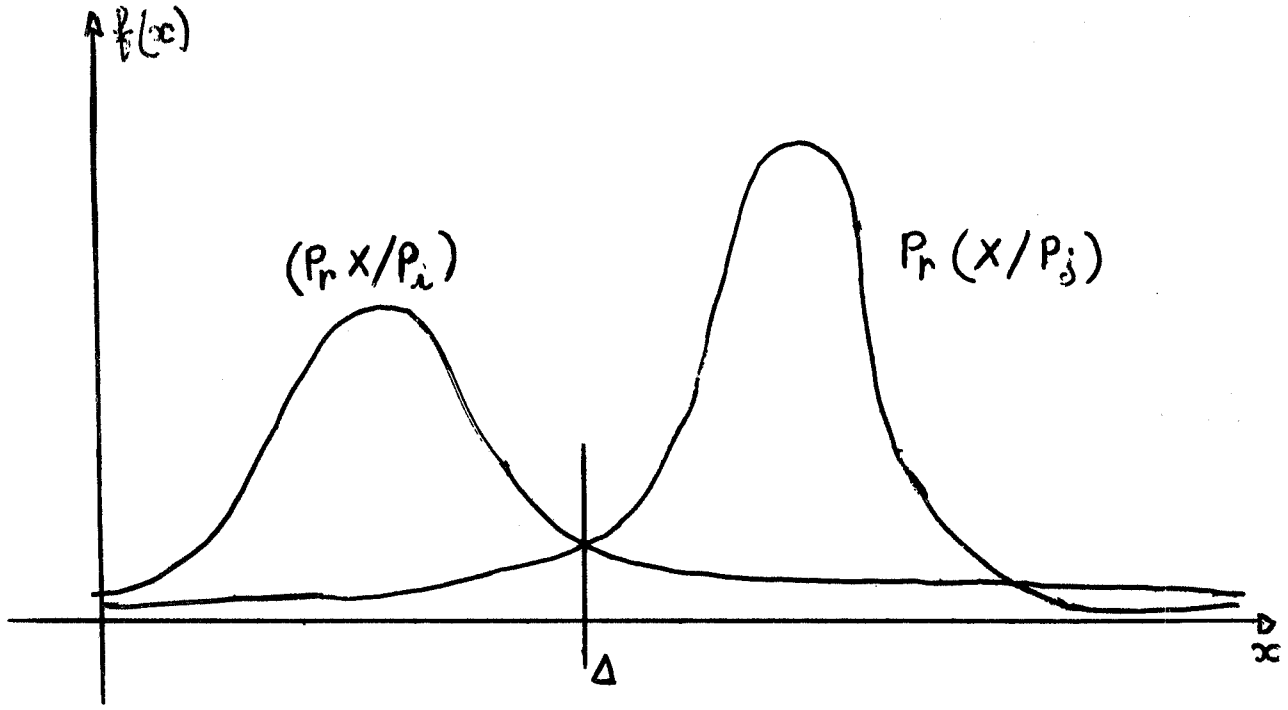


figure 6 : Séparation des phonèmes P_i et P_j à l'aide de l'amplitude X qu'ils développent sur un canal du vocodeur.

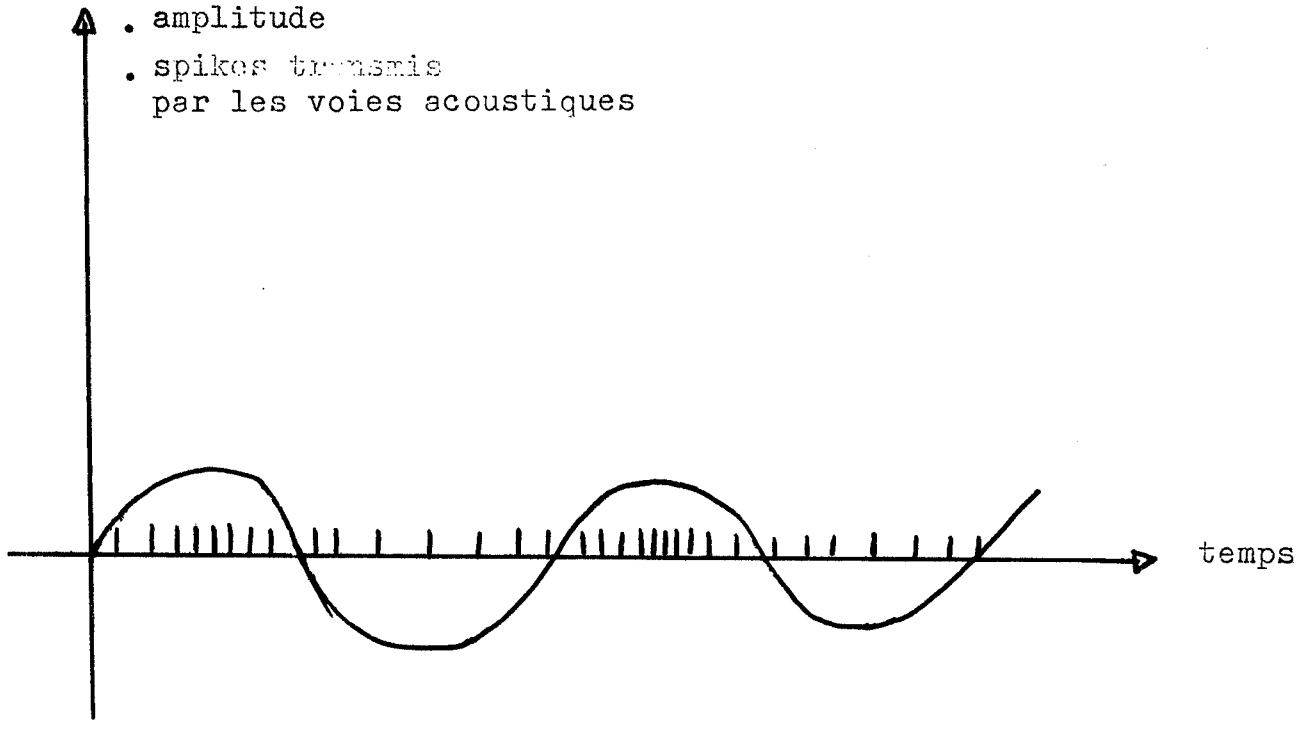


figure 7 : Codage d'une amplitude par l'organe de Corti : la fréquence de récurence des impulsions transmises par les voies acoustiques obéit à une modulation de fréquence.

BIBLIOGRAPHIE

[1] N.ABRAMSON - Information Theory and Coding (1962) Mac Graw New York

[2] D.L.RICHARDS - Distorsion of speech by quantizing. Electronic letter (juin 1967) Vol.3 - N° 6

[3] C.BERGER-VACHON et G.MESNARD - Efficacité d'un système de codage. Applications à la parole (à paraître)

[4] G.FERRIEU, J.PONCIN, G.ROUX, J.VINCENT-CARREFOUR - Synthèse et reconnaissance de la parole par calculateur - L'Echo des recherches - juin 1968

[5] J.L.FLANAGAN - Speech analysis - Synthesis and perception - Second Expanded Edition (1972) Springer Verlag - Berlin

[6] E.AKERMAN - Biophysical science (paragraphe 6-2 : neural mecanism of hearing) Prentice Hall Inc. Englewood Cliff - N.J. (1962) pp. 105-117.

UN MODELE MATHEMATIQUE DE COCHLEE

J. CAELEN*- G. PERENNOU**

* Assistant I.U.T. Toulouse

* * Professeur I.U.T. Toulouse

RESUME :

Partant d'un modèle mathématique de l'oreille, se présentant sous forme d'un système d'équations différentielles non linéaires, on décrit le comportement de la membrane basilaire pour diverses valeurs des paramètres mécaniques et de propagation et pour divers signaux d'entrée.

Des représentations sont obtenues par résolution numérique du système de départ. On essaie ensuite d'envisager la recherche des formants à partir des informations disponibles sur la membrane. Les résultats sont comparés à ceux obtenus par l'étude directe du signal.

ABSTRACT :

From a mathematical model of the ear, presented like a non linear differential equation system, we shall describe the behaviour of the basilar membrane for several values of mechanics and propagation parameters and for several signals.

Representation are obtained by numerical resolution of the equation system. Then, we shall try to research the formants from available information on the membrane. Results are confronted to those obtained by direct analysis of signal.

UN MODELE MATHEMATIQUE DE COCHLEE

J. CAELEN - G. PERENNOU

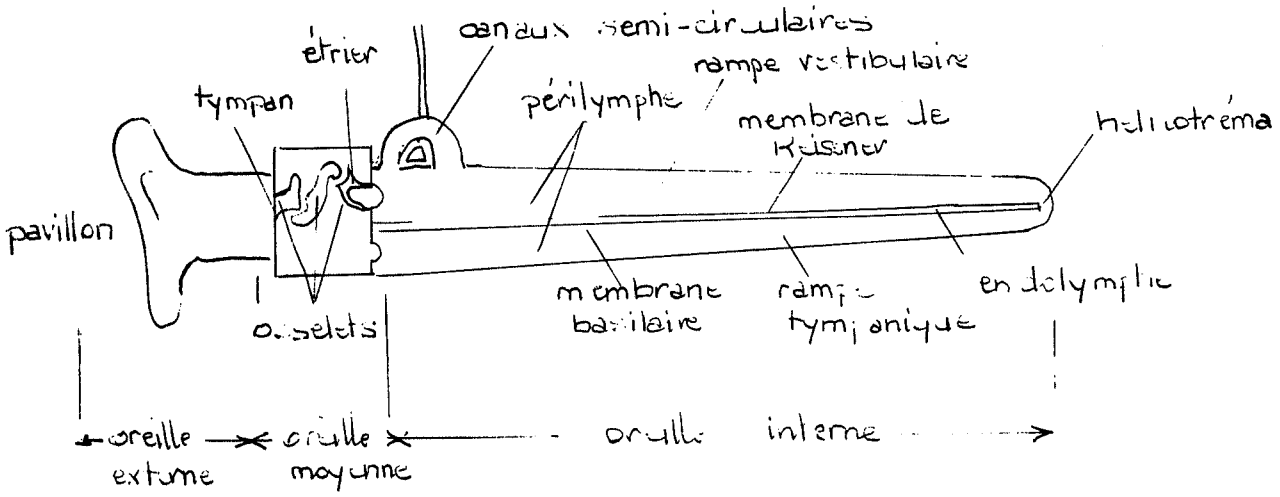
1. INTRODUCTION

Dans le but de mieux étudier le signal-parole et d'en préciser l'analyse nous avons mis en oeuvre un modèle mathématique de cochlée fondé sur des observations connues et des hypothèses généralement admises. (Voir Rapport interne du C.E.R.F.I.A. Février 1974).

Différents modèles ont été proposés par Békésy, Flanagan, Fievet, Maissis, Walrave et Alinat par exemple. Notre propos sera de donner quelques résultats obtenus par notre modèle et de poser quelques questions à propos du fonctionnement du système auditif cellulaire en rapport avec l'information disponible sur la membrane basilaire.

2. DESCRIPTION RAPIDE DU MODELE

Soit la cochlée développée :



Le système ossiculaire transmet les vibrations sonores aériennes au canal cochléaire en adaptant l'impédance des deux milieux . Une onde de compression est propagée à l'intérieur de la cochlée, dans la périlymphe, liquide interne, à travers les rampes vestibulaires et tympaniques qui communiquent à l' hélicotrema .

D'aprés Békésy cette onde peut être considérée comme se propageant dans une seule direction. Sous l'effet des différences de pression dans les deux rampes provoquées lors de la propagation de l'onde, la membrane basilaire entre en vibration en se déformant transversalement. Cette membrane a des propriétés mécaniques intéressantes, sa masse croit vers l'hélicotréma tandis que sa rigidité diminue. Békésy remarque que dans ces conditions, pour des signaux sinusoidaux, elle présente des maxima de vibration localisés en fonction de la fréquence.

EQUATION DE L'ONDE DANS LA PERILYMPHE :

L'idée essentielle est de considérer la propagation dans la périlymphe comme peu perturbée par les vibrations de la membrane basilaire dans une région suffisamment éloignée de cette membrane.

Si nous supposons la propagation adiabatique l'équation de propagation s'écrit :

$$\frac{\partial^2 \varphi}{\partial t^2} = \frac{c_0^2}{\left(1 + \frac{\partial \varphi}{\partial x}\right)^{\gamma+1}} \cdot \frac{\partial^2 \varphi}{\partial x^2} \qquad \text{(Beyer)}$$

avec C_0 célérité à l'origine

φ déplacement des particules par rapport à leur position à l'équilibre x

γ constante du fluide

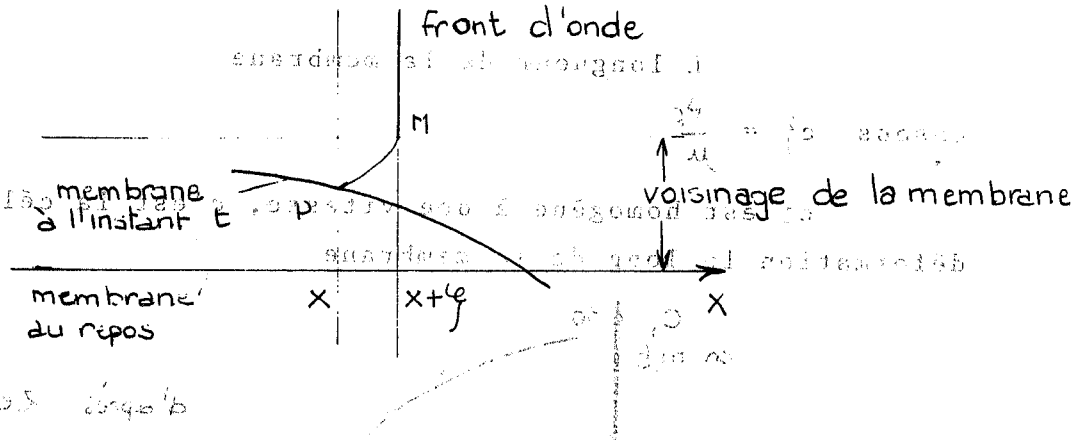
et les conditions initiales $u(x,0) = 0$ en posant $u(x,t) = \frac{\partial \varphi}{\partial t}$

et aux limites $u(0,t) = \varphi(t)$

$\varphi(t)$ fonction excitation donnée à l'origine

on montre que la viscosité, les distosions harmoniques et les effets de bord sont négligeables.

Au voisinage de la membrane :



Le long du front d'onde la pression est constante si bien que la valeur calculée au point M dans le fluide est la même qu'au point P sur la membrane donc :

$$P_p(x) = P_M(x+\xi)$$

On peut voir que le voisinage de la membrane est le siège de tourbillons puisque $\mathrm{rot}(u) \neq 0$

EQUATION DE LA MEMBRANE VIBRANTE :

On écrit l'équation de la dynamique sur une portion de membrane réduite à une seule dimension en corrigeant par un terme tenant compte de la tension transversale. Nous obtenons :

$$\mu \frac{\partial^2 h}{\partial t^2} = P_f \frac{\partial^2 h}{\partial x^2} + \gamma \frac{\partial h}{\partial t} + \lambda h = P - P'$$

P pression dans la rampe vestibulaire

P' pression dans la rampe tympanique

avec μ densité de masse au point x

P_f coefficient de rigidité longitudinale au point x

λ coefficient de rigidité transversale au point x

γ amortissement visqueux

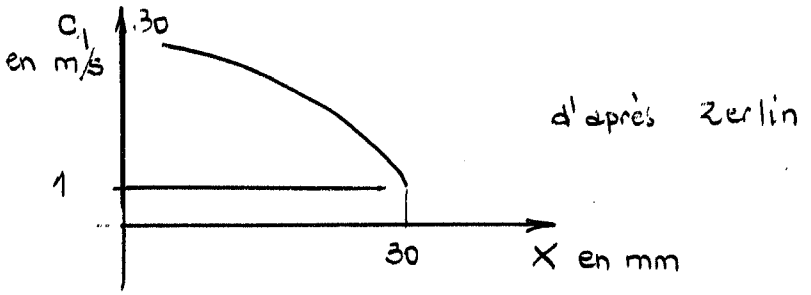
et les conditions initiales $h(x,0) = \frac{\partial h}{\partial t}(x,0) = 0$

et aux limites $h(0,t) = h(L,t) = 0$ L longueur de la membrane

L longueur de la membrane

posons $c_1^2 = \frac{P_f}{\mu}$

c_1 est homogène à une vitesse, c'est la célérité de la déformation le long de la membrane



Le système résumant les principaux phénomènes de vibration est :

$$(1) \quad \frac{\partial^2 \varphi}{\partial t^2} = \frac{c_0^2}{\left(1 + \frac{\partial \varphi}{\partial x}\right)} \gamma_{11} \frac{\partial^2 \varphi}{\partial x^2}$$

$$(2) \quad P(x+\varphi, t) = \rho_0 c_0 u(x+\varphi, t)$$

$$(3) \quad P'(2L-x-\varphi, t) = \rho_0 c_0 u(2L-x-\varphi, t)$$

$$(4) \quad \frac{1}{c_1^2(x)} \frac{\partial^2 h}{\partial t^2} - \frac{\partial^2 h}{\partial x^2} + f(x) \frac{\partial h}{\partial t} + \lambda(x) h = \frac{P - P'}{P_f(x)}$$

Avec les conditions initiales et aux limites :

$$u(x,0) = 0 \qquad u(0,t) = \varphi(t)$$
$$h(x,0) = 0 \qquad h(0,t) = h(L,t) = 0$$
$$\frac{\partial h}{\partial t}(x,0) = 0$$

RESOLUTION NUMERIQUE :

On montre que la déformation de l'onde est négligeable au cours de la courte propagation (2 x 30mn). La vitesse de phase est pratiquement constante le milieu est faiblement dispersif. Les distorsions harmoniques sont de l'ordre de 10^{-5} par rapport à la fréquence fondamentale. Une approximation est faite pour les fréquences audibles qui sont des fréquences moyennes.

On prendra :

$$u(x,t) = u_0 e^{-\alpha x} \varphi\left(t - \frac{x}{c}\right) \qquad c \leq c_0$$

$\varphi(t)$ excitation à l'origine

L'équation (4) est linéaire à coefficients variables. Une méthode de transformation de Fourier n'est pas applicable, on la résoud par un schéma aux différences finies. On choisira un schéma implicite qui a l'avantage d'être toujours stable.

Si on appelle Δx et Δt les pas de discrétisation, alors y_{ij} sera une approximation de $h(i\Delta x, j\Delta t)$.

L'équation s'écrira :

$$\frac{y_i^{j+1} - 2y_i^j + y_i^{j-1}}{(c_i \Delta t)^2} - \frac{(y_{i+1}^{j+1} - 2y_i^{j+1} + y_{i-1}^{j+1}) + \varepsilon(y_{i+1}^j - 2y_i^j + y_{i-1}^j) + (y_{i+1}^{j-1} - 2y_i^{j-1} + y_{i-1}^{j-1})}{4(\Delta x)^2}$$

$$+ \lambda \frac{y_i^j - y_i^{j-1}}{\Delta t} + \gamma y_i^j = \beta \varphi_{ij} - \beta' \varphi'_{ij}$$

β et β' coefficients d'amortissement

Avec les valeurs initiales :

$$y_i^0 = y_i^1 = 0$$

les valeurs aux limites :

$$y_0^j = y_n^j = 0$$

3. QUELQUES RESULTATS SUR LA SIMULATION

Un jeu d'essai des paramètres a été utilisé pour étudier le comportement du modèle à diverses stimulations.

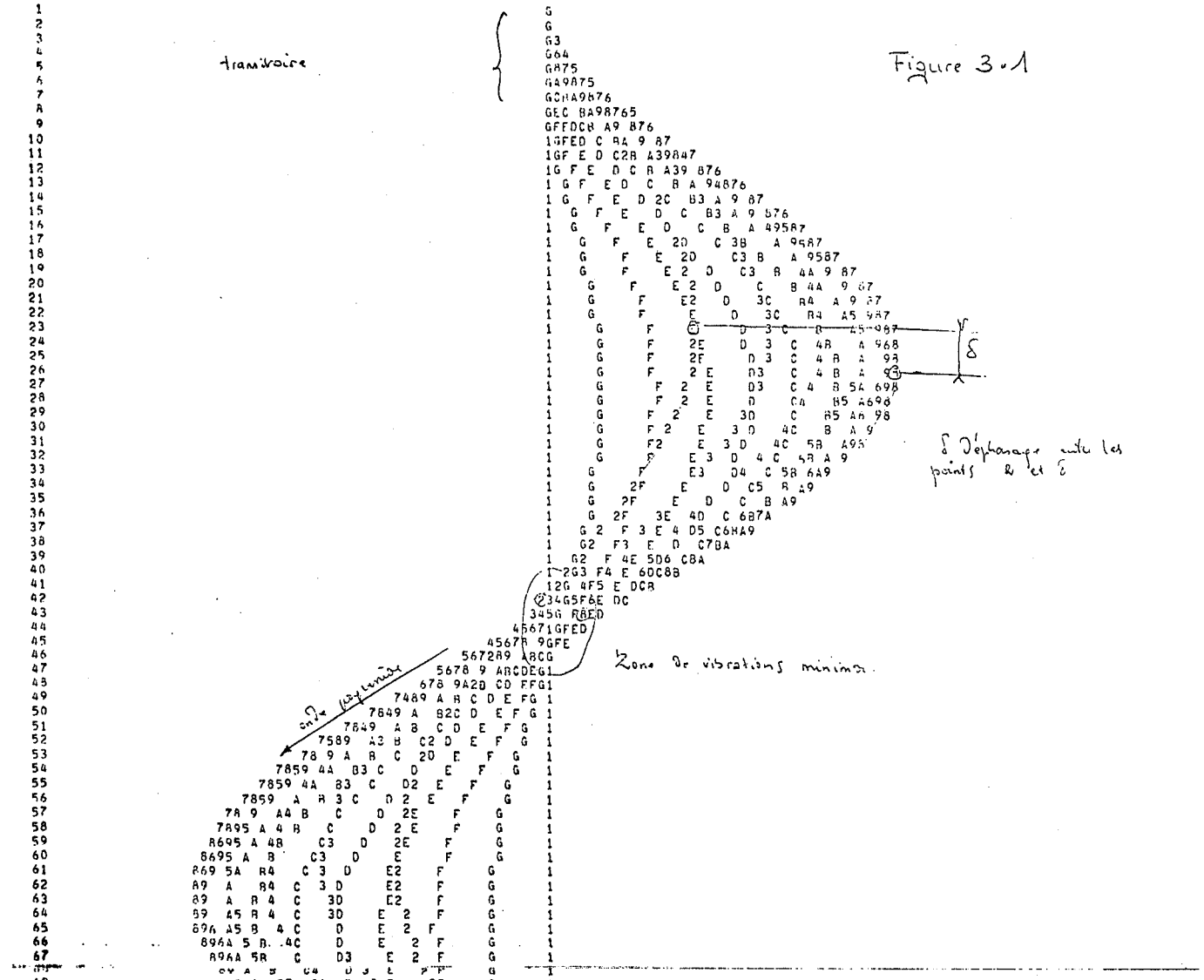
De façon générale on distingue un maximum de vibration localisé qui dépend de la fréquence pour des sons sinusoidaux purs mais aussi des coefficients d'amortissement de la membrane et de l'onde propagée dans le liquide puisque c'est elle qui induit la vibration dans la membrane basilaire. En particulier elle crée des déphasages entre les divers points de la membrane. Ces déphasages diminuent avec la célérité de l'onde dans le fluide du moins pour les points de la membrane proche de l'étrier. Pour les autres, au contraire apparaît une distorsion de phase. Le maximum de vibration se déplace également vers l'étrier si, l'amortissement de l'onde augmente. A la limite si l'onde est très amortie ce maximum est très près de l'étrier et la membrane se comporte comme un filtre passe-bas et inversement si l'onde n'est pas amortie la membrane se comporte seulement comme une ligne à retard sans filtrage. Les déformations propagées de la membrane dépendent évidemment de ses propres caractéristiques. En fonction des coefficients d'élasticité transversales et longitudinales on peut constater des distorsions vers l'extrémité apicale.

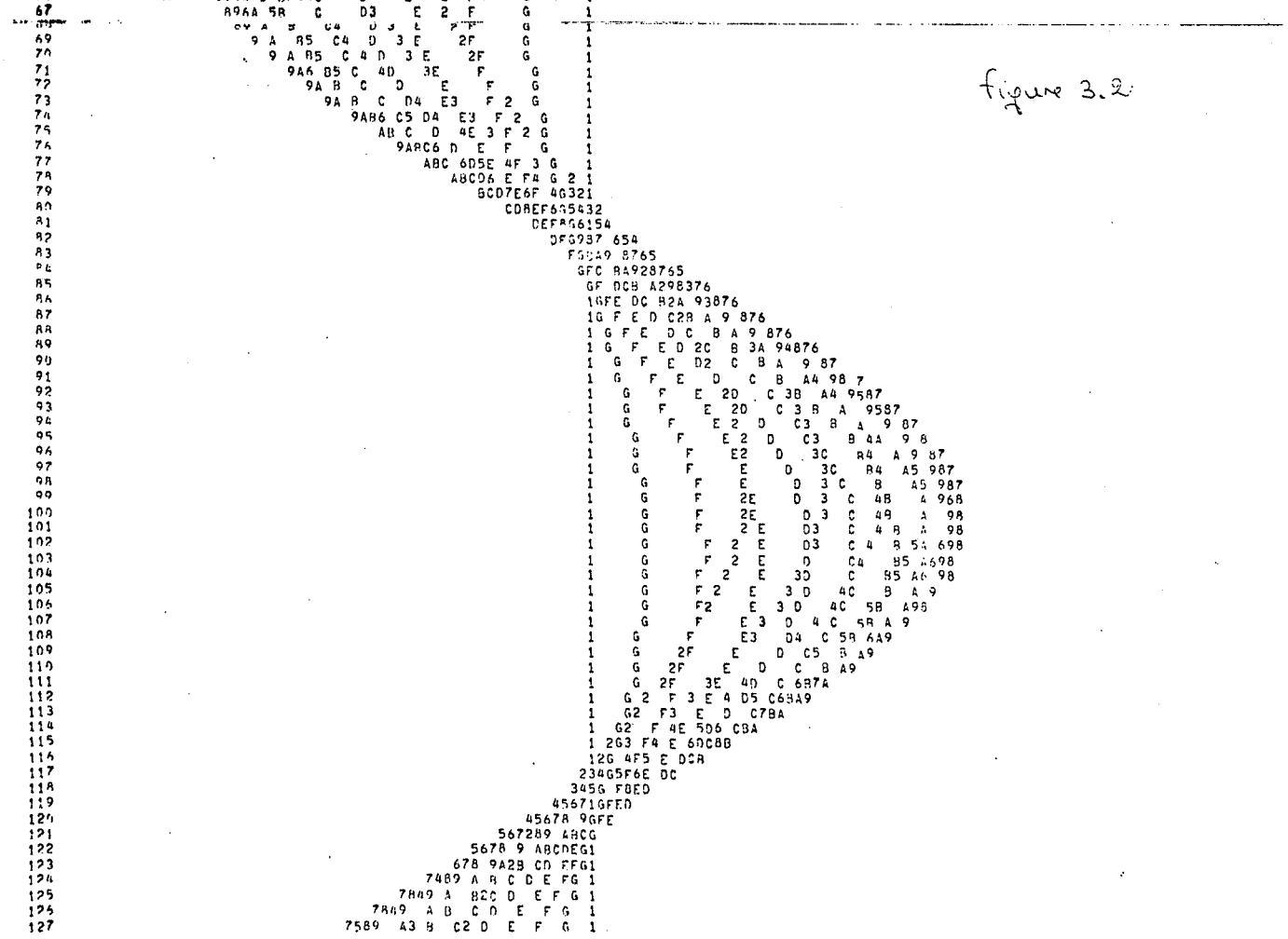
Pour des sons complexes tels ceux de la parole il est difficile de parler de localisation des fréquences. On peut seulement remarquer que les points proches de l'étrier suivent fidèlement les variations du signal mais que plus on s'éloigne plus les vibrations suivent de loin ces variations. On constate un maximum de vibration mais il est difficile de dire qu'il dépend de la fréquence.

Une analyse du signal porté par la membrane est faite par des méthodes d'analyse développées par MM. Dours et Facca (voir exposé)

Sur les figures ci-après on peut voir les vibrations point par point de la membrane. En joignant les caractères 1 on peut suivre les vibrations du premier point de la membrane (près de l'étrier) au cours du temps. En joignant les caractères 2, puis 3 ... G on peut suivre les fonctions des points 2,3 .... 17. L'ensemble des courbes ainsi obtenues constitue la réponse à un ton pur de fréquence 600 hz. On constate un maximum d'amplitude au point 8.

CELERITE C=60.0 SIGNAL DE FREQUENCE 200.





4. ANALYSE DES SIGNAUX PAR DES METHODES DE TYPE PERIODOGRAMME

Soit un signal $f(t) = a_1 \sin(\omega_1 t + \varphi_1) + a_2 \sin(\omega_2 t + \varphi_2)$

On généralisera facilement au cas de plusieurs sinusoïdes.

Soit $\{ t_1 \dots t_n \dots \}$ un ensemble de discrétisation Δt constant

On pose

$$f_n^*(t) = \frac{1}{n} \sum_{i=0}^{n-1} f(t - i\Delta t)$$

$f_n^*(t)$ est la convolution de f(t) par la fenêtre rectangulaire $\frac{1}{n}$ de longueur $n\Delta t$

Par la formule des arcs d'accroissement constant on obtient

$$f_n^*(t) = a_1 \frac{\sin\left(\omega_1 t - \frac{n-1}{2}\omega_1 \Delta t\right) \sin \frac{n\omega_1 \Delta t}{2}}{n \sin \frac{\omega_1 \Delta t}{2}} + a_2 \frac{\sin\left(\omega_2 t - \frac{n-1}{2}\omega_2 \Delta t\right) \sin \frac{n\omega_2 \Delta t}{2}}{n \sin \frac{\omega_2 \Delta t}{2}}$$

Pour une valeur de n_1 proche de la période : $\frac{2\pi}{\omega_1 \Delta t} = T_1$

soit $n_1 = \frac{2\pi}{\omega_1 \Delta t} + \varepsilon$

nous avons :

$$f_{n_1}^*(t) \simeq a_1 \frac{\varepsilon x_1^2}{\pi \sin x_1} \sin(\omega_1 t + x_1 + \varphi_1) + a_2 \frac{x_1}{\pi} \frac{\sin \frac{\pi\omega_2}{\omega_1}}{\sin x_2} \sin\left(\omega_2 t + x_2 + \varphi_2 - \frac{\pi\omega_1}{\omega_2}\right)$$

en posant

$$x_1 = \frac{\omega_1 \Delta t}{2} = \pi \nu_1 \Delta t$$

$$x_2 = \frac{\omega_2 \Delta t}{2} = \pi \nu_2 \Delta t$$

avec $\nu_1 \leq \frac{1}{2\Delta t}$ et $\nu_2 \leq \frac{1}{2\Delta t}$ d'après le théorème de Shannon.

Donc

$$0 \leq x_1 \leq \frac{\pi}{2}$$

$$0 \leq x_2 \leq \frac{\pi}{2}$$

Si on appelle :

$$f_n^{*p}(t) = \frac{1}{n} \sum_{i=0}^{n-1} f^{*p-1}(t - i\Delta t)$$

alors on voit que :

$$f_{n_1}^{*p} = a_1 \left(\frac{c x_1^2}{\pi \sin x_1} \right)^p \sin(\omega_1 t + p x_1 + \varphi_1) + a_2 \frac{x_1}{\pi} \left(\frac{\sin \frac{\pi \omega_2}{\omega_1}}{\sin x_2} \right)^p \sin\left(\omega_2 t + p x_2 + \varphi_2 - p \frac{\pi \omega_2}{\omega_1} \right)$$

posons $$r_{n_1} = a_1 \left(\frac{c x_1^2}{\pi \sin x_1} \right)^p \sin(\omega_1 t + p x_1 + \varphi_1)$$

alors $$\lim_{p \to \infty} r_{n_1} = 0$$

et $$\lim_{p \to \infty} f_{n_1}^{*p} = K \sin(\omega_2 t + \varphi)$$

avec $$K = \lim_{p \to \infty} a_1 \left(\frac{x_1}{\pi} \frac{\sin \frac{\pi \omega_2}{\omega_1}}{\sin x_2} \right)^p$$

On peut affirmer de même qu'il existe n_2 proche de $\frac{2\pi}{\omega_2 \Delta t}$ tel que

$$\lim_{q \to \infty} \left(f_{n_1}^{*p} \right)_{n_2}^{*q} = 0$$

<u>Méthode</u> : On cherche les valeurs n_1, n_2 qui donnent à p fixé le résidu quadratique minimum, n_1 approche donc la période d'une composante du signal. On recommence avec n_2 pour trouver la deuxième composante la plus prépondérante et ainsi de suite. Par différences avec le signal origine on extrait toutes les composantes ainsi éliminées.

Une méthode, de sommation locale, pourrait suggérer une tentative d'explication du fonctionnement des cellules auditives. Il est en effet admis que des cellules externes effectuent par leur potentiel de sommation, une "moyenne" sur lespotentiels microphoniques des cellules internes. Ces deux potentiels déclenchent ensuite dans le nerf auditif l'influx nerveux ou potentiel d'action, suite d'impulsions d'égale amplitude. Ici encore on peut penser au

théorème d'échantillonnage : "La moyenne d'une fonction aléatoire est donnée par le nombre d'échantillons écrétés par une fonction seuil uniformément répartie sur le support de la fonction aléatoire."[11]

Ces considérations permettraient peut être d'envisager une méthode d'extraction des formants en rapport avec un fonctionnement possible des cellules auditives.

5. CONCLUSION

Nous avons donné quelques résultats concernant le transfert mécanique du signal sonore et avancé quelques idées quant au transfert nerveux de ce signal. L'objectif de cet exposé n'est pas de donner une théorie du système auditif mais plutôt d'inspirer quelques réflexions sur le problème de l'analyse du signal et de poser quelques questions sur le fonctionnement encore mal connu du codage nerveux.

B I B L I O G R A P H I E

- - - - - - - - - - - - - - - -

(1) Von Békésy Experiments in hearing (1960)
(2) L Pimonuw Vibrations en régime transitoire (1962)
(3) J.J Matras Acoustique et électroacoustique (Eyrolles)
(4) J.S Lienard Thèse sur l'analyse et la synthèse de la parole
(5) Y. Rocard Dynamique générale des vibrations (1971) (Masson)
(6) Matthieu et Fleury Vibrations mécaniques, acoustique (Eyrolles)
(7) Flanagan Analysis Synthésis and Parception
(8) Fievet, Maissis, Walrave la reconnaissance en temps réél de la parole Automatisme (1970)
(9) Wesley le Mars Nyborg Acoustic Streaming
(10) Uno Nigul Plane stress waves in membranes caused by an arbitrary pressure wave (JASA 51 n°1 1972)
(11) M. Bouix Propagation sur une demi-droite (revue CETHEDEC 1971)
(12) R.T. Beyer Acoustique non-linéaire (Pysical acoustics 1965)
(13) Strange Ross Transmission de l'oreille moyenne (JASA 43 n°3 1968)
(14) Francis Hugh Fenlon An extension of the Bessel-Fubini series for a multiple frequency CW acoustic source of finite amplitude (JASA 51 n°1 1972)
(15) Some properties of longitudinal Shear Waves (JASA 43 n°5 1968)
(16) Measuring the parameter of nonlinearity of liquids, L.E Hargrove & K. Achyuthan
(17) R.T Beyer Propagation d'un signal sonore transitoire dans un liquide visqueux (JASA 44 n°2 1968)
(18) Richard F. Salant Acoustic wave propagation part a sinusoidal surface (JASA 44 n°1 1968)
(19) Fubini-Girhon Alta frequenza 4.530 (1935)
(20) D.T Blackstock (JASA 34 1962)
(21) Stanley Zerlin (JASA 46 n° 4 1969)
(22) G. Valiron Equations fonctionnelles et applications (Masson)
(23) R.D Richtmyer Méthode des différences finies pour les problèmes de conditions initiales
(24) Alinat Une cochlée artificielle (Journées d'études du Galf)
(25) Atal et hanauer
(26) Gueguen Le filtrage optimal de Kalmann

ANALYSE TEMPORELLE DU SIGNAL VOCAL COMPAREE A L'ANALYSE FREQUENTIELLE CLASSIQUE DU POINT DE VUE DE LA RECONNAISSANCE

-:-:-:-

D. DOURS * - R. FACCA * - G. PERENNOU **

-:-:-:-

C. E. R. F. I. A. Toulouse

Résumé :

On se propose de comparer l'efficacité discriminante des paramètres obtenus d'une part, par une méthode classique d'analyse spectrale du signal vocal, d'autre part, par une méthode d'analyse temporelle du signal en synchronisme avec le fondamental.

A cette fin, une méthode de Reconnaissance des formes est mise en oeuvre permettant la comparaison des résultats obtenus sur les deux ensembles de paramètres extraits d'un même échantillon de voyelles, associées ou non à des consonnes, issues de plusieurs locuteurs.

Abstract

We intend to compare the parameters discriminant efficiency, on the one hand through a classical method of speech spectral analysis, and on the other hand through a pitch synchronous time domain speech analysis.

For this purpose, we are using a pattern recognition method, allowing the comparison betwen the results given by the two sets of parameters obtained from a same sample of vowels, associated or not to consonants, produced by different speakers.

* Assistant I.U.T. Toulouse

** Professeur I.U.T. Toulouse

ANALYSE TEMPORELLE DU SIGNAL PAROLE
COMPAREE A
L'ANALYSE FREQUENTIELLE DU POINT DE VUE DE LA RECONNAISSANCE

-:-:-:-:-:-:-

D. DOURS - R. FACCA - G. PERENNOU

-:-:-:-:-:-:-

I - INTRODUCTION

Cette étude s'insère dans le projet de reconnaissance de la parole étudié dans notre laboratoire, projet qui s'articule conformément au schéma de la figure 1.

Dans le cadre de ce projet, les études antérieures de reconnaissance se basaient sur des paramètres obtenus à partir d'un vocoder*. Ce sont les réflexions sur les performances ainsi obtenues, qui nous ont amené à mettre l'accent sur la recherche de paramètres physiques par un traitement direct du signal vocal, et plus particulièrement sur ceux des voyelles prélevées dans un mot.

L'objet de cette communication est de présenter les premiers résultats obtenus dans cette direction.

* Fournit par le C.N.E.T. de Lannion. Contrat D.R.M.E.

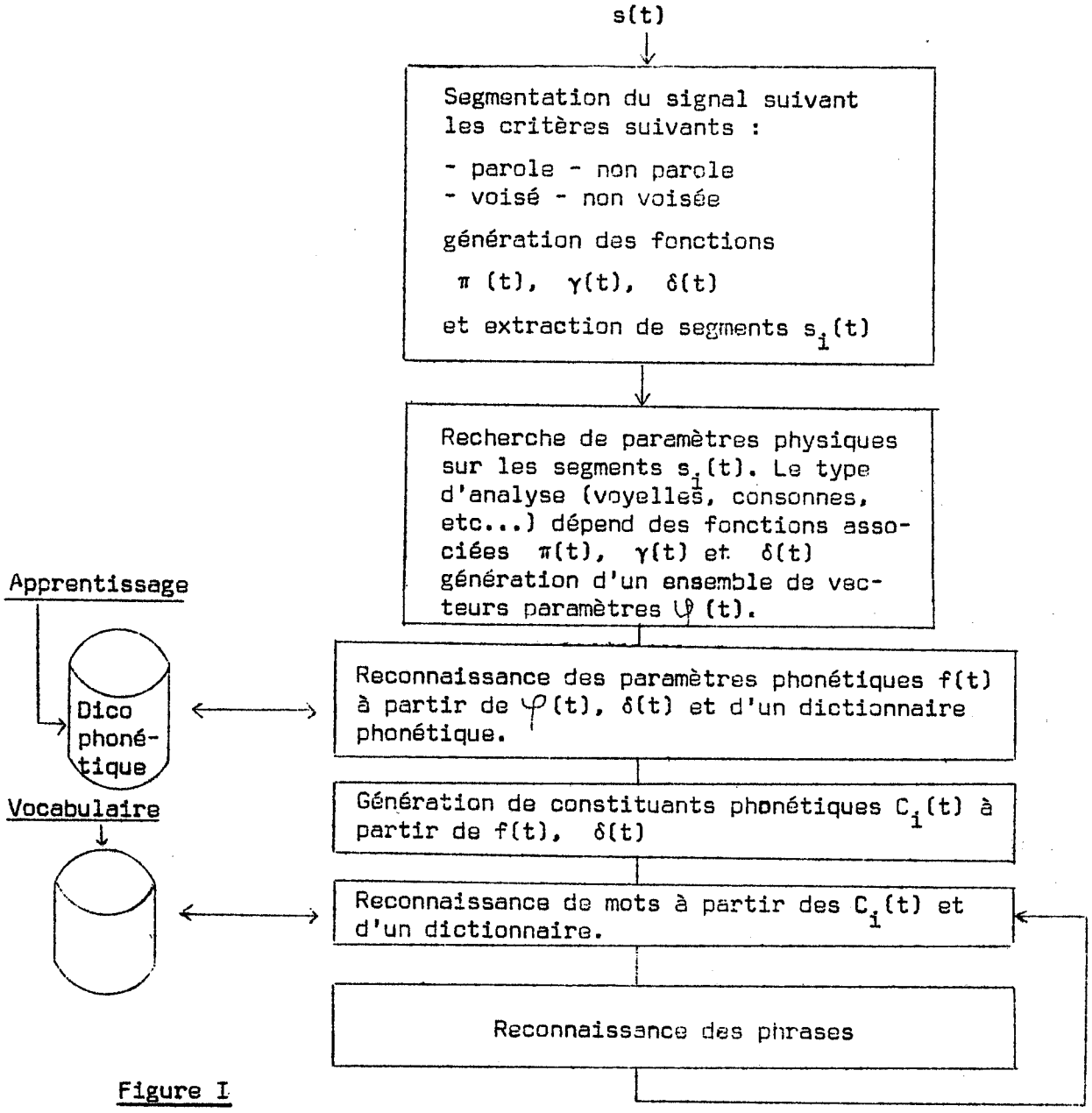


Figure I

$$\pi(t) = \begin{cases} 1 & \text{si l'instant t est le début d'une période de fondamental} \\ 0 & \text{sinon} \end{cases}$$

$$\gamma(t) = \begin{cases} 1 & \text{aux instants } t_i,\ t_i + \Delta t, \ldots, t_i + n\Delta t \text{ sur les intervalles de signal parole non voisés.} \\ 0 & \text{ailleurs} \end{cases}$$

$$\delta(t) = \begin{cases} 1 & \text{si le début d'une zone de signal parole} \\ 1 & \text{si apparition d'une zone de signal non parole.} \end{cases}$$

II - METHODE D'ANALYSE DIRECTE DU SIGNAL PARLE

2.1. Motivations de l'étude :

Les méthodes du type analyse spectrale en dépit de leurs avantages, notamment la simplicité de mise en oeuvre et rapidité d'analyse (vocoder) présentent un certain nombre d'inconvénients bien connus lorsqu'il s'agit du signal parlé. En effet, la modulation du spectre par le fondamental, la définition d'un spectre à court terme, le caractère convolutionnel de la production de la parole, font que ces méthodes ne donnent pas toujours les résultas escomptés surtout en ce qui concerne la précision des paramètres obtenus.

Diverses méthodes ont été proposées pour pallier ces inconvénients. Elles sont en général fondées sur l'analyse directe du signal. Citons principalement les méthodes d'analyse par codage prédictif introduites par Atal-Schroeder-Hanauer (3), (9),l'analyse par filtrage optimal de Kalman proposée par Guegen et Carayannis (6), la méthode présentée par ITAKURA et SAITO (12), (13).

La méthode d'analyse que nous avons envisagée s'apparente à celle proposée par E.N. Pinson (10)par rapport à laquelle elle introduit des éléments nouveaux, notamment la détection du pitch et l'analyse séparée des formants.

2.2. Modèle du système de phonation et relations entre les différentes méthodes :

Les méthodes d'analyse directe du signal dont nous venons de parler utilisent toutes le modèle proposé par Fant (4).

Dans ce modèle on considère le signal vocal comme la réponse s(t) d'un système acoustique f(t) (conduit vocal, conduit nasal, radiation buccale) à une excitation e(t) (séries d'impulsions quasi-périodiques provenant des cordes vocales dans le cas des sons voisés et d'un bruit blanc dû aux turbulences de l'air dans les constrictions du canal vocal dans le cas des sons non voisés).

Conformément à ce modèle, la fonction glottale e(t) peut être définie de la façon suivante :

Une période du fondamental de durée T peut être divisée en deux intervalles T_f et T_0 pendant lesquels la glotte est respectivement fermée et ouverte.

$$e(t) = g(t) \qquad t \in T_0$$

$$e(t) = 0 \qquad t \in T_f$$

Le modèle étant linéaire et stationnaire pendant une période du fondamental, la réponde du système durant l'intervalle T_f est entièrement déterminée par les conditions initiales au début de l'intervalle non excité et par la réponse libre du système f(t).

Il est facile de vérifier que sous les hypothèses classique (canal-vocal assimilé à une série de résonnateurs en cascade), la réponse libre du système est de la forme

$$f(t) = \sum_{i=1}^{N} e^{-\alpha_i t} \; (a_i \sin \omega_i t + b_i \cos \omega_i t)$$

avec :

$\omega_i = 2 \pi F_i$ (F_i : fréquence du $i^{ème}$ formant)

$\alpha_i = \pi B_i$ (α_i : coefficient d'amortissement du $i^{ème}$ formant)

a_i et b_i caractérisent l'amplitude de la fréquence F_i.

N est le nombre de formants. Il dépend du modèle choisi (3 ou 4).

Ce modèle suggère diverses méthodes d'analyse ayant toutes pour but l'obtention d'un ensemble de paramètres caractérisant le signal vocal.

Ces paramètres pourront être obtenus :

1°/ Par identification de la fonction de transfert (transformée en Z) du système de phonation sur chaque période du fondamental.

Parmi ces méthodes conduisant à la mise en oeuvre de filtres linéaires numériques, on peut citer :

a) l'analyse par codage prédictif introduite par Atal-Hanauer, dans laquelle le critère à optimiser est un critère d'écart quadratique entre le signal réel et le signal estimé.

b) L'analyse par filtrage optimal de Kalman introduite par Guegen-Carayannis qui fournit une estimation optimale des paramètres au sens du minimum de variance.

c) L'analyse présentée par ITAKURA-SAITO dans laquelle le critère considéré est un critère de maximum de vraisemblance.

Ces méthodes qui ne diffèrent que par le processus de résolution adopté, permettent si on le désire, de déterminer les pôles de la fonctions de transfert à partir des paramètres obtenus.

2°/ Par identification de la réponse impulsionnelle du système de phonation sur chaque période du fondamental.

La méthode d'analyse que nous développons est une méthode directe en synchronisme avec le Pitch. Elle s'apparente à celle proposée par E.N.Pinson en ce sens qu'elle consiste à approximer directement le signal dans des tranches de temps pendant lesquelles le signal glottal est négligeable.

2.3. Principe de la méthode proposée :

Nous avons vu que lorsque la glotte est fermée, la réponse libre du système à laquelle, pour tenir compte des conditions d'obtention du signal, nous avons ajouté une composante continue, est de la forme :

$$f(t) = \sum_{i=1}^{N} e^{-\pi B_i t} \left(a_i \sin 2\pi F_i t + b_i \cos 2\pi F_i t\right) + a_0$$

Soit $s(t) = f(t) + B(t)$ le signal observé pendant que la glotte est fermée, dans lequel B(t) est un bruit stationnaire indépendant du signal.

Le problème consiste à identifier les paramètres du modèle $(a_0, a_i, b_i, F_i, B_i)$ $i = 1, N$ de telle manière qu'un critère déterminé soit minimisé. Nous avons choisi l'erreur quadratique moyenne :

$$E = \int_{T_f} [s(t) - f(t)]^2 \; h^2(t)\, dt \qquad \text{(h(t) fonction de pondération).}$$

2.3.1. Principe général d'approximation :

Soit l'ensemble des paramètres du modèle $\{a_0, a_i, b_i, F_i, B_i / i=1,N\}$. On constate que pour (B_i, F_i) $i=1,N$ fixés, f(t) est une combinaison linéaire des paramètres (a_0, a_i, b_i) $i=1,N$ pour chaque valeur de t.

On considère alors les deux ensembles :

$$\{a_0, a_i, b_i / i=1,N\} \quad \text{et} \quad \{B_i, F_i / i=1,N\}.$$

1°/ Pour $\{B_i, F_i / i=1,N\}$ fixé, on cherche

$\{a_0, a_i, b_i / i=1,N\}$ qui minimise l'erreur E.

Cela revient à résoudre le système :

$$\frac{\partial E}{\partial a_0} = 0 \quad \frac{\partial E}{\partial a_i} = 0 \quad \frac{\partial E}{\partial b_i} = 0 \qquad i=1,N$$

Comme f(t) est une combinaison linéaire des paramètres (a_0, a_i, b_i) i=1,N, le système sera résolu par une méthode matricielle.

2°/ En fixant les paramètres (a_0, a_i, b_i) i=1,N aux valeurs trouvées au 1°/, on recherche alors $\{B_i, F_i / i=1,N\}$ qui minimise l'erreur E.

Il faut donc résoudre le système :

$$\frac{\partial E}{\partial F_i} = 0 \quad \frac{\partial E}{\partial B_i} = 0 \qquad i = 1,N$$

f (t) n'étant pas linéaire par rapport aux paramètres (B_i, F_i) i=1,N, le système sera résolu par une méthode du gradient. Nous avons choisi une méthode à pas séparés.

A partir des paramètres (B_i, F_i) i=1,N ainsi obtenus on iterre le processus tant que l'erreur E décroit.

2.3.2. Méthode d'approximation séparée des formants

On se propose de rechercher successivement les différents formants.

Conformément à la théorie développée dans le chapitre (2.2.), nous faisons l'hypothèse que, lorsque la glotte est fermée, le système oscille librement et qu'il est de la forme :

$$f(t) = \sum_{i=1}^{N} f_i(t) \quad \text{avec } f_i(t) = A_i e^{-\pi B_i t} \sin(2\pi F_i t + \varphi_i)$$

Soit s(t) = f(t) + B(t) le signal observé, B(t) étant un bruit stationnaire indépendant du signal, et soit $\hat{s}(t)$ le modèle du signal à un seul formant.

$$\hat{s}(t) = Ae^{-\pi Bt} \sin(2\pi Ft + \varphi)$$

Nous cherchons à identifier les paramètres (A,φ,B,F) du modèle de telle manière qu'un critère déterminé soit minimisé :

nous avons choisi l'erreur quadratique moyenne :

$$E = \int_{T_f} [\hat{s}(t) - s(t)]^2 \, dt$$

Nous donnerons tout d'abord une justification théorique, puis la mise en oeuvre pratique de la méthode.

Justification théorique

La recherche séparée des formants peut se justifier par les considérations théoriques suivantes.

L'expression de l'erreur E peut se développer sous la forme :

$$E = \int_{T_f} \hat{s}^2(t) \, dt - \sum_{i=1}^{N} \int_{T_f} \hat{s}(t) \, f_i(t) \, dt + cte$$

Nous avons regroupé dans la constante tous les termes du développement indépendants de $\hat{s}(t)$. En effet ces termes ne jouent aucun rôle dans la minimisation de l'erreur quadratique.

Tous calculs faits, et en posant $T_f = [0,T]$ l'expression de E devient :

$$E = \frac{A^2}{4\pi} \left[\frac{X}{B} - \frac{Y}{2(B^2+4F^2)} \right] - \sum_{i=1}^{N} \frac{AA_i}{\pi} \left[\frac{X_i}{((B+B_i)^2 + (F-F_i)^2)} - \right.$$

$$\left. - \frac{Y_i}{(B+B_i)^2 + (F+F_i)^2} \right]$$

en posant : $\Theta = 2\pi FT + \varphi$

$$\Theta'_i = 2\pi T (F-F_i) + \varphi - \varphi_i \qquad \varphi'_i = \varphi - \varphi_i$$

$$\Theta''_i = 2\pi T (F+F_i) + \varphi + \varphi_i \qquad \varphi''_i = \varphi + \varphi_i$$

$$X = 1-e^{-2\pi BT}$$

$$Y = e^{-2\pi BT}\,[-B\cos 2\,\theta + 2\,F\sin 2\,\theta]-[-B\cos 2\varphi + 2\,F\sin 2\varphi]$$

$$X_i = e^{-\pi BT}\,[(B+B_i)\cos\theta_i' + 2(F-F_i)\sin\theta_i']-[-(B+B_i)\cos\varphi_i' + 2\,(F-F_i)\sin\varphi_i']$$

$$Y_i = e^{-\pi BT}[-(B+B_i)\cos\theta_i'' + 2\,(F+F_i)\sin\theta_i'']-[-(B+B_i)\cos\varphi_i'' + 2\,(F-F_i)\sin\varphi_i'']$$

En examinant l'expression de E, on constate que :

lorsque $F \to F_i$ l'expression sous le signe somme se comporte comme $\frac{AA_j}{\pi}\,\frac{X_j}{(B+B_j)^2}$, les autres termes devenant négligeables devant celui-là.

Il s'en suit que lors de la minimisation de l'erreur quadratique, ce terme sera prépondérant vis-à-vis des autres et ceci d'autant plus que l'amplitude A_j du jème formant est grande devant l'amplitude des autres formants.

De même lors des dérivations par rapport à chaque paramètre du modèle les constations vont dans le même sens.

En première approximation, tout se passe pour le gradient comme si le signal ne comportait qu'un seul formant.

Mise en oeuvre pratique

Généralement, le premier formant possède plus d'énergie que le second qui en a plus que le troisième. Mais ceci n'est pas toujours le cas chez certains locuteurs en particulier pour la voyelle i. Comme les formants sont obtenus successivement en fonction de l'énergie qu'ils portent, nous allons effectuer une recherche guidée par une méthode tenant compte de ce phénomène.

Soit $\{A, \varphi, B, F\}$ l'ensemble des paramètres du modèle. En posant

$X = \{A, \varphi\}$ et $Y = \{B, F\}$, le modèle du signal à un seul formant s'écrit

$$\hat{s}(X,Y,t) = Ae^{-\pi Bt}\sin(2\pi Ft + \varphi)$$

On se donne un certain nombre de valeurs Y_i de départ :

$$Y_i = (F_i, B) \quad \begin{cases} F_i \in \Omega = \{300, 550, \ldots, 3400\} \\ B = 50 \end{cases}$$

et pour chaque Y_i on met en oeuvre l'algorithme d'approximation séparée des formants sur le signal s(t).

A l'issue de cette recherche systématique, on examine le résidu quadratique le plus faible.

Soient $\hat{X}_1$, $\hat{Y}_1$ les paramètres correspondants et soit $\hat{s}(\hat{X}_1, \hat{Y}_1, t)$ l'approximation du formant prépondérant.

On introduit alors un nouveau signal,

$$s_1(t) = s(t) - \hat{s}(\hat{X}_1, \hat{Y}_1, t)$$

Le même processus est appliqué à $s_1(t)$ pour une plage de fréquences $\Omega_1 = \Omega - V_1$,

avec V_1 ensemble de fréquences de Ω les plus voisines de $\hat{F}_1$ (dans notre cas nous prenons 4 fréquences).

La réduction de Ω à Ω_1 peut être faite car on sait que les fréquences des autres formants ne seront pas dans cette plage. On gagne ainsi en temps de calcul.

Pour extraire le formant suivant, on introduit un nouveau signal $s_2(t)$ et on applique le proceseus pour une plage de fréquence $\Omega_2=\Omega_1 - V_2$.

Il faut remarquer que cette méthode permet d'obtenir des paramètres assez précis et présente l'avantage de n'avoir besoin d'aucune connaissance du signal.

2.3.3. Mise en oeuvre pratique de la méthode

Lors de l'analyse du signal vocal, la détermination du fondamental et plus particulièrement le début de chaque période se révèle très importante.

Extraction de la fréquence fondamentale

Le début de chaque période du fondamental est caractérisé par une importante et rapide variation d'énergie. Cependant, en se basant uniquement sur ce critère, il apparaît souvent de fausse détections. C'est pourquoi, parallèlement, par une méthode de périodogramme, l'existence d'une fréquence F_0 est recherchée. Lorsqu'on aboutit à l'existence de cette fréquence et connaissant sa valeur, on élimine pratiquement toutes les fausses détections en respectant les deux principes suivants :

1°/ Dans une tranche de signal ne prendre en considération que les variations d'énergie les plus importantes,

2°/ Ne retenir que celles qui sont adaptées à la fréquence F_0.

Ceci nous permet d'obtenir dans de bonnes conditions la fonction π (t) suivante :

π (t) = 1 si l'instant t est le début d'une période du fondamental

π (t) = 0 sinon

Il est alors possible de déterminer une succession d'intervalles de temps où le signal est quasi périodique.

Revenons maintenant au cas où aucune périodicité de type F_0 n'est mise en évidence.

On utilise alors la fonction $\gamma(t)$ définie par :

$\gamma(t) = 1$ aux instants t_i, $t_i+\Delta t,\ldots,t_i+n\Delta t$ sur les intervalles de signal parole non voisés

$\gamma(t) = 0$ ailleurs

Le temps t_i est déterminé par le premier échec dans la détermination d'une fréquence F_0 et le découpage du signal est effectué à l'aide de la fonction $\gamma(t)$.

En résume, si le signal est voisé, le découpage est effectué en synchronisme avec le fondamental, s'il n'est pas voisé, il est arbitrairement découpé en intervalles de temps de durée Δt donnée à l'avance.

Détermination des paramètres

Dans le cas d'un signal voisé, nous mettons en oeuvre la méthode générale d'approximation. Elle nécessite toutefois une connaissance approximative des paramètres (B_i et F_i) i=1,N.

En effet, la fonction $E = \int_{Tf} (s(t) - f(t))^2 \, dt$ n'est pas convexe.

Il faut donc se donner une valeur initiale des paramètres proche de la solution, sinon on risque de converger vers des maxima locaux qui ne correspondent pas au problème à résoudre, et d'autre part, on augmente le temps de calcul inutilement.

Pour obtenir l'approximation des paramètres (B_i, F_i) i=1,N, nous commençons par mettre en oeuvre la méthode des formants séparés, car elle ne nécessite aucune connaissance sur le signal.

La méthode s'articule donc ainsi :

1. Initialisation du processus :

Sur la première tranche de signal, nous appliquons la méthode des formants séparés. Ceci nous permet d'obtenir $\{B_i, F_i / i=1,N\}$ avec une bonne précision et d'initialiser la méthode globale sur les 3 formants pour affiner les résultats.

2. Processus de suivie :

On suppose que le signal évolue peu d'une tranche de signal à l'autre. On utilise alors les paramètres $\{B_i, F_i / i=1,N\}$ obtenus sur la tranche de signal précédente pour initialiser la méthode globale sur les 3 formants.

Si l'algorithme converge, on itère sur le processus de suivie. Sinon c'est que le signal évolue rapidement. Dans ce cas, on itère sur le processus d'initialisation.

Ceci permet de générer la fonction vectorielle $\varphi(t)$ des paramètres physiques du signal.

III - CONCLUSION

La méthode présentée permet de caractériser le signal vocal par un ensemble réduit de paramètres qui peuvent être favorablement comparés du point de vue de la reconnaissance, à ceux obtenus par d'autres méthodes et en particulier les méthodes spectrales.

Le fait que cette méthode ne présente aucun des inconvénients des méthodes spectrales, influe directement sur la qualité des paramètres obtenus.

Il faut remarquer toutefois que les paramètres obtenus contiennent aussi des informations relatives aux locuteurs ce qui nécessite la mise en oeuvre d'une méthode de normalisation en vue d'une utilisation ultérieure pour la reconnaissance.

BIBLIOGRAPHIE

-:-:-:-:-

1. L. RABINER, R. SCHAFER, C. RADER : The Chirp transform Algorithm and its Application. Bell System Technical Journal, June 1969.

2. M. NOLL : Cepstrum Pitch Determination. Journal of the Acoustical Society of America, Vol 41, pp 293-309, 1967.

3. B. ATAL, S. HANAUER : Speech Analysis and Synthesis by linear Prediction of the Speech Wave. Journal of the Acoustical Society of America, Vol 50, pages 637-655, 1971.

4. G. FANT : Acoustic theory of speech production. Mouton the Hague. Paris 1970.

5. J. FLANAGAN : Speech Analysis Synthesis and Perception. Second Edition, Springer - Verlag Berlin, Heidelberg-New-York, 1972.

6. C.J. GUEGEN, G. CARAYANNIS : Analyse de la parole par filtrage optimal de KALMAN, 1973.

7. A. OPPENHEIM : Speech Analysis-Synthesis System Based on Homomorphic Filtering. Journal of the Accustical Society of America, Vol 45 pages 458-465, 1969.

8. R. SCHAFER, L. RABINER : System for Automatic Formant Analysis of Voicet Speech. Journal of the Acoustical Society of America, Vol 47, Pages 634-648, 1970.

9. B. S. ATAL and SCHROEDER : Adaptative Predictive Coding of Speech Signals. Bell system Tech. J. 49, pp 1973-1986, 1970.

10. E.N. PINSON : Pitch-Synchronous Time-Domain Estimation of Formant Frequencies and Bandwiths. Journal of The Acoustical Society of America, Vol 35 Pages 1265-1273, 1963.

11. A. H. MAISSIS : Une méthode d'extraction du fondamental. L'ONDE ELECTRIQUE, Vol. 53, Fasc. 3, Pages 110-112, 1973.

12. F. ITAKURA - S. SAITO : A statistical method for estimation of speech spectral density and formant frequencies.
Electronics and communications in Japan Vol. 53 A n° 1 1970.

13. F. ITAKURA - S. SAITO : An analysis-synthesis telephony based on maximum likelihood method.
Proc. Int. Congr. Accoust. C. 5. 5. Tokyo- 1968.

METHODE DES TRAJECTOIRES :

UNE MESURE DE DISTANCE ENTRE LIGNES POLYGONALES ORIENTEES

L. F. PAU

Laboratoire d'Automatique
Ecole Nationale Supérieure des Télécommunications

46 rue BARRAULT
75634 - PARIS CEDEX 13

Résumé

On rappelle le lien existant entre les processus stochastiques multidimensionnels tels que la voix codée non segmentée et leur représentation sous forme de trajectoires orientées. Aux fins de la segmentation par les transitions, ou de la reconnaissance globale de trajectoires tels que les mots, on définit une mesure de distance entre trajectoires polygonales orientées dans R^n, comptant des nombres de sommets différents.

Abstract

One is reminded of the existing relationship between multidimensional stochastic processes such as coded and unsegmented speech, and their visualisation as oriented trajectories. For the purpose of segmentation on the basis of transitions, or for global recognition of trajectories such as spoken words, we define a metric applicable to polygonal oriented trajectories in R^n, having each different numbers of summits.

METHODE DES TRAJECTOIRES :

UNE MESURE DE DISTANCE ENTRE LIGNES POLYGONALES ORIENTEES

L. F. PAU

1. POSITION DU PROBLEME ET METHODE DES TRAJECTOIRES

On rappelle qu'il est toujours possible de représenter un processus stochastique multidimensionnel dans R^n par une trajectoire dans R^n. Par trajectoire, nous entendons courbe orientée dépendant d'un seul paramètre réel, ici le temps t. Dans bien des problèmes pratiques, l'entier n sera néanmoins la dimension de l'espace réduit, obtenu après compression de processus d'apprentissage de dimension supérieure. Dans [3], il est ainsi démontré comment, comprimant la sortie échantillonnée d'un vocoder à p canaux, $p > n$, l'analyse factorielle des correspondances conduisait à représenter sur une carte factorielle ($n = 2$) des trajectoires représentant chacune un mot analysé par le vocoder. De cette manière, on ouvrait la voie à une nouvelle représentation visuelle de la parole non segmentée par des trajectoires, complétant ainsi les images complètes que sont les sonogrammes.

Il devient alors indispensable de pouvoir comparer entre elles deux trajectoires orientées quelconques dans R^n, en conformité avec la perception humaine. Ceci doit être fait au moyen d'une mesure de similarité, ou mieux une métrique, conduisant aux applications suivantes :

a) approximation d'une trajectoire par une trajectoire plus simple, telle une ligne polygonale à peu de sommets : ceci constitue un procédé de segmentation de la parole qui, au lieu de chercher les "zones stables" s'attachera à limiter spatialement les transitions entre "zones stables" (par exemple : phonèmes ou"segments minimaux") ;

b) reconnaissance globale d'un mot, par comparaison entre la trajectoire mesurée, et des trajectoires d'apprentissage ; ceci a donné [3] un taux de reconnaissance de l'ordre de 80 % dans un échantillon de 215 mots constituant aussi les mots d'apprentissage.

Si tant est que le problème ait été abordé, on a utilisé par ailleurs les concepts de similarité suivants :

α) plus courte perpendiculaire commune aux deux trajectoires ;

β) valeur moyenne de la distance euclidienne entre points des deux trajectoires.

Ces concepts ont les inconvénients suivants :

1) α) est difficile à calculer, défini de manière multivoque, et utilise une propriété exclusivement locale (fig. 1) ;

2) ni α) ni β) ne tiennent compte des orientations relatives des deux trajectoires ; dans la fig. 2, les mots T1, T2 d'une part, et T1, T3 d'autre part seront identifiés par α) et β), ce qui est absurde ;

3) ni α) ni β) ne tiennent compte des rebroussements relatifs des deux trajectoires (fig. 3).

2. MESURE DE DISTANCE ENTRE LIGNES POLYGONALES ORIENTEES

Dans ce paragraphe, nous esquissons une explication de la théorie introduite dans [2] et détaillée dans [1] :

21. Donnée sous forme discrète des trajectoires orientées

$T1 = (x1_i \ , \ i = 1 \ , \ n(T1) \ , \ x1_i \ \varepsilon \ R^n)$

$T2 = (x2_i \ , \ i = 1 \ , \ n(T2) \ , \ x2_i \ \varepsilon \ R^n)$

où l'orientation est celle fixée par l'ordre des points.

22. Recherche de la trajectoire partielle

la plus proche de l'autre trajectoire, successivement pour T1 et T2 ; la trajectoire partielle P_{T2} (T1) de T1 la plus proche de T2 est constituée par la suite ordonnée des points de T1 les plus proches des points de T2 prélevés dans l'ordre de ces derniers ; P_{T2} (T1) comptera donc autant de points que T2 (fig. 4). On se donne donc une métrique dans R^n.

23. *Voisinage convexe commun V(T1, T2)*

C'est l'intersection des deux domaines de R^n, obtenus comme compromis convexes des points de T1 $\cup$ $P_{T1}(T2)$ et T2 $\cup$ $P_{T2}(T1)$ respectivement (fig. 5).

24. *Orientation d'une trajectoire par rapport à une autre*

L'opérateur $P_{\bullet}(\bullet)$ transmet dans la trajectoire-argument l'orientation de la trajectoire-indice. On calculera, et on notera θ (T1, T2), le nombre de rebroussements de $P_{T2}(T1) \cap V(T1, T2)$ par rapport à $T1 \cap V(T1,T2)$. Deux trajectoires sont de même sens ssi θ (T1, T2) = 0 (fig. 4).

25. *Aire gauche A (T1, T2) séparant les trajectoires T1, T2*

Si T1, T2 ne sont ni de même sens, ni de sens contraire, on opèrera sur ces trajectoires une anti-inversion locale aux points de rebroussement localisés précédemment. Cette opération aura pour but de rectifier localement l'orientation relative de T1 et T2. Si T1, T2 sont soit de même sens, soit de sens contraire, après rectifications éventuelles, on pourra calculer l'aire gauche les séparant en sommant les aires des triangles ou quadrilatères définis par les suites ordonnées de sommets sur $P_{T2}(T1) \cap V(T1, T2)$ $P_{T1}(T2) \cap V(T1, T2)$ -(fig 6).

26. *Distance d (T1, T2) entre les trajectoires orientées T1, T2*

$$d(T1,T2) = f(\theta(T1,T2)) \times \frac{2 \times A(T1,T2)}{L(P_{T2}(T1) \cap V) + L(P_{T1}(T2) \cap V)}$$

où

- *f est une fonction monotone croissante possédant certaines propriétés ;*
- *L (•) est la longueur d'une trajectoire, obtenue par intégration de l'abcisse curviligne.*

d est donc, de manière simplifiée, égale au quotient pénalisé de l'aire séparant T1, T2 dans le voisinage commun V, par la longueur moyenne de T1, T2 dans celui-ci.

3. CONCLUSION

La recherche d'une distance entre courbes orientées dans R^n semble répondre à une préoccupation fondamentale en reconnaissance des formes dynamiques, et plus généralement aux fins de comparer ou classer des processus aléatoires multidimensionnels. La métrique proposée semble, sur la base des premiers tests, être calculable en des temps courts.

4. REFERENCES

(1) *L. F. PAU : Mesure de distance entre lignes polygonales orientées ; applications en reconnaissance des formes - Laboratoire d'Automatique, ENST, Paris, Mars 1974 ; contient toutes les références complémentaires*

(2) *L. F. PAU : Common theoretical formulation of the pattern recognition, identification and detection problems, IMSOR, Technical Univ. Denmark, Copenhague, Mars 1972, 24 p.*

(3) *L. F. PAU : Statistical reduction and recognition of speech patterns, in Machine perception of patterns and pictures, Institute of Physics Conf. Publ. 13, London, 1972*

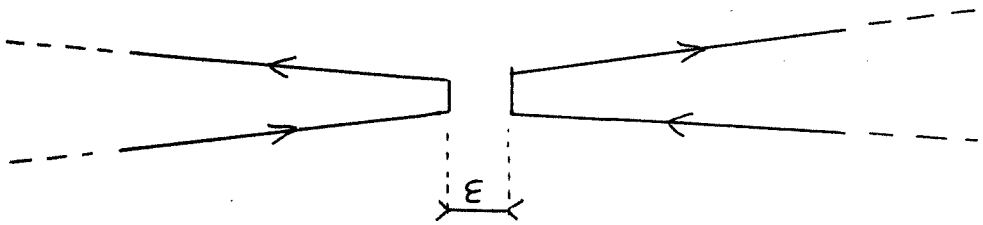


Figure 1 : Distance ε-infiniment petite au sens de ∝)

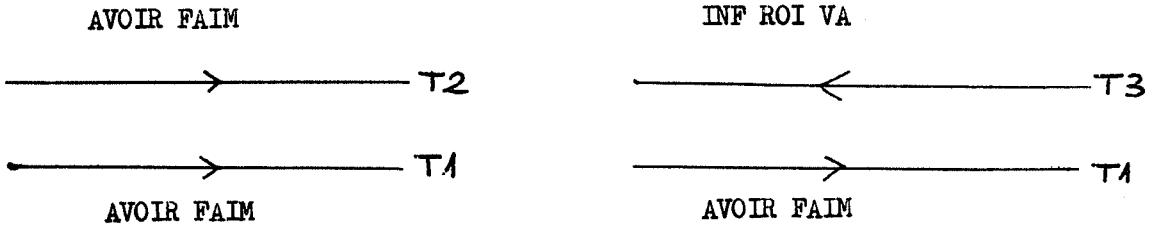


Figure 2 : Comparaison de phrases inversées

Figure 3 : La distance entre t1 et t2 devrait dépendre du nombre de radians dont a tourné la flèche extrémité de t2

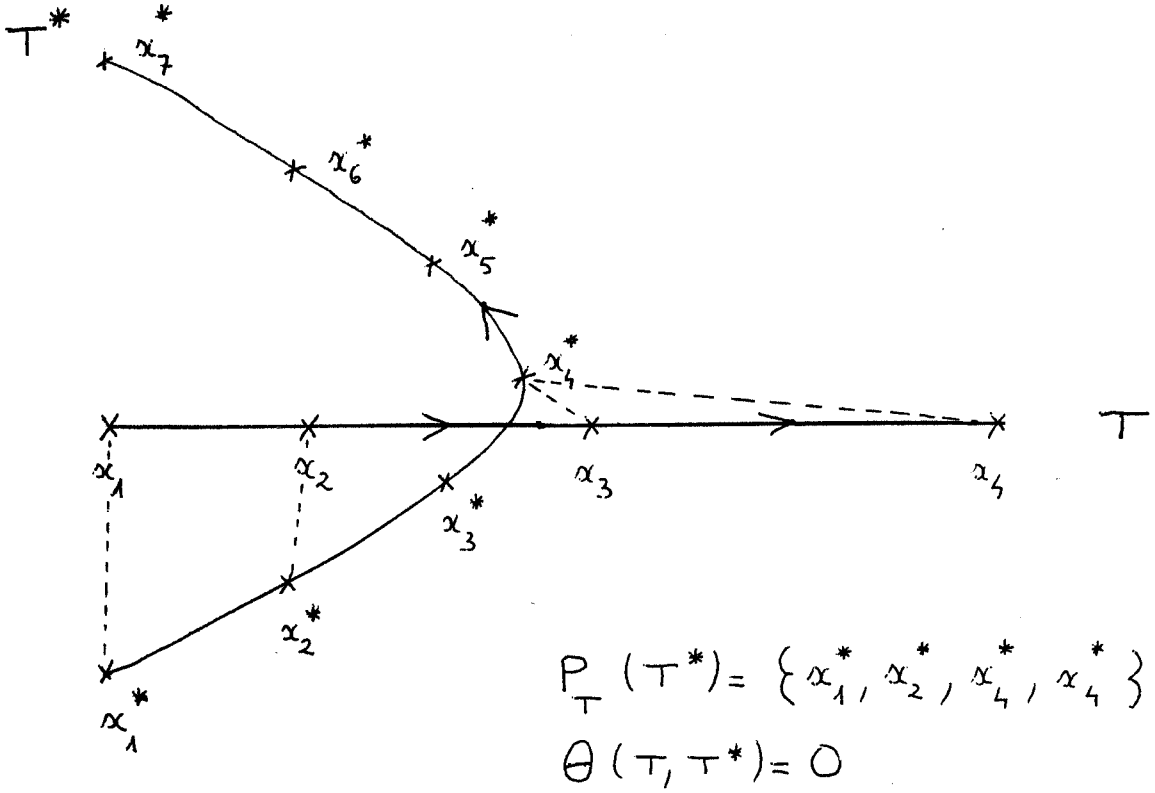


Figure 4 : Trajectoire partielle de T la plus proche de T .

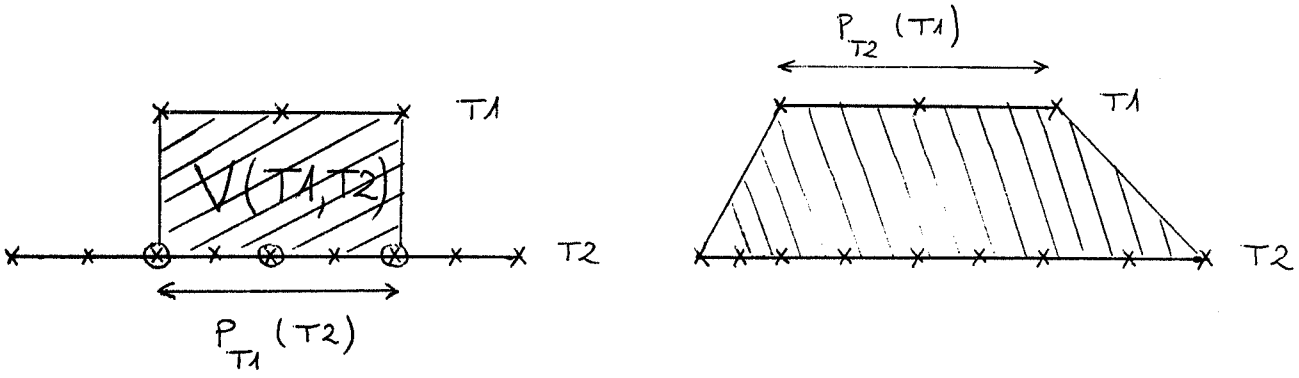


Figure 5 : Voisinage convexe commun V(T1,T2) : il figure ici à gauche par construction

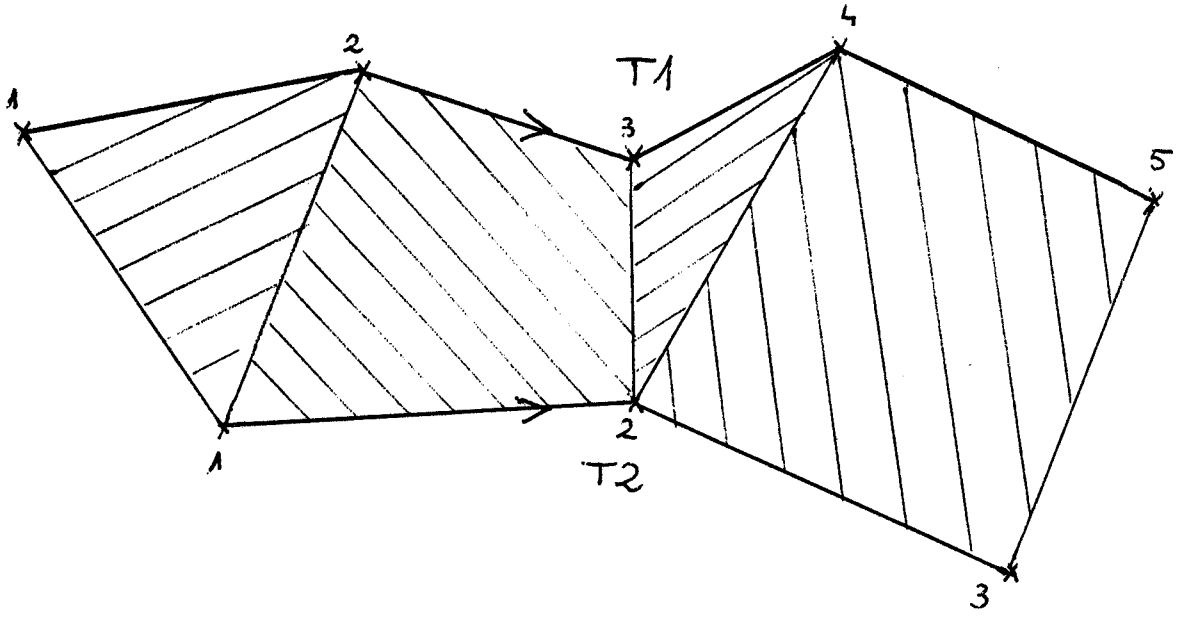


Figure 6 : Aire gauche A(T1,T2)

APPLICATION DE L'ANALYSE LINEAIRE DISCRIMINANTE A LA RECONNAISSANCE

J.J. MASSOT

LABORATOIRES DE MARCOUSSIS
DEPARTEMENT RECHERCHES PHYSIQUES DE BASE
Groupe Traitement du Signal

RESUME

L'analyse discriminante linéaire est une technique de séparation d'évènements en classes. Elle consiste à chercher des axes de l'espace de représentation sur lesquels les classes se séparent au mieux et conduit à chercher les directions propres du rapport de deux matrices de co-variance. Elle a été appliquée avec succès à la reconnaissance des voyelles du français.

APPLICATION OF LINEAR DISCRIMINANT ANALYSIS IN RECOGNITION

ABSTRACT

Linear discriminant analysis is a technique for separating events in classes. It consists of looking for axis of space on which classes are separated the best. For that, we compute eigenvectors of the quotient of two covariant matrix. The method has been successfully applied in recognition of french vowels.

APPLICATIONS DE L'ANALYSE DISCRIMINANTE LINEAIRE EN RECONNAISSANCE

I - INTRODUCTION .-

L'analyse discriminante linéaire est une technique générale de séparation d'évènements en classes.

Elle consiste à chercher des axes d'un espace vectoriel de représentation des évènements sur lesquels les projections des points représentatifs des évènements d'une même classe soient le mieux regroupés possible (au sens de leur variance).

Cette technique a été appliquée à la reconnaissance des voyelles.

II - ETUDE THEORIQUE .-

On se donne des individus repérés chacun par un ensemble de mesures, et que l'on sait affecter a priori dans des classes. On se propose de chercher des axes sur lesquels les projections des individus appartenant à une même classe soient les plus regroupées possible.

La Figure III montre de tels axes. Chaque classe est constituée par un nuage de points de forme sensiblement elliptique dans un espace $(\vec{e}_1, \vec{e}_2)$ à deux dimensions. La figure montre bien que la distance euclidienne dans le ré-

férentiel est insuffisante pour classer des points tels que A $(\overline{G_1A} \quad \overline{G_2A})$ alors que les projections de ces points sur un axe tel que $\vec{u}$ fournissent la solution.

Mathématiquement, on cherche un axe qui remplisse les conditions suivantes simultanément :

1/ - pour chaque classe la variance des projections des points de la classe est minimale,

2/ - par l'ensemble des classes, la variance des projections des centres de gravité des classes est maximale.

La condition 1/ s'écrit : soit V_c la matrice de co-variance des points d'une classe, u une direction quelconque de l'espace, il faut écrire que la quantité (1) $\vec{u}^T V_c \vec{u}$ est minimale.

Cette condition doit être étendue à toutes les classes. Comme les quantités telles que (1) sont positives, il suffit d'écrire que leur somme est minimale, soit :

$$\vec{u}^T \left(\Sigma V_c\right) \vec{u} \quad \text{minimum.}$$

La condition 2/ s'écrit, de même, si B est la matrice de co-variance des centres de gravité :

$$\vec{u}^T B \vec{u} \quad \text{maximum}$$

La suite du calcul relève de la théorie des valeurs propres. Un problème équivalent aux deux conditions est de maximiser le rapport des deux quantités, soit :

$$\frac{\vec{u}^T B \vec{u}}{\vec{u}^T W \vec{u}} \quad \text{en posant : } W = \Sigma V_c$$

Une simplification supplémentaire apparaît si l'on remarque que

$$T = W + B \text{ (théorème de Huyghens)}$$

où T est la matrice de co-variance totale du nuage.

Les solutions sont alors les solutions de l'équation aux valeurs propres

$$T^{-1} B \vec{u} = \lambda \vec{u}$$

équation où l'on prendra la direction attachée à la plus grande valeur propre.

III - APPLICATIONS A LA RECONNAISSANCE DES VOYELLES .-

Un fichier a été constitué, contenant diverses élocutions des voyelles du français. L'espace de base est constitué par les 16 raies du spectre calculé par FFT sur 32 points successifs, échantillonnage à 8 kHz.

Huit classes de voyelles ont été considérées dans l'expérience : A (pâte) E (oeuf) I, O (beau), U (lu) ET (été) AI (lait, OU (mou). 7 élocutions de chaque voyelle sont utilisées, dites par différents locuteurs. Le programme écrit calcule à partir de ces données les directions discriminantes, et les projections des centres de gravité des classes sur ces directions. Dans l'application exposée ici, il suffit de retenir les trois directions attachées aux trois plus fortes valeurs propres pour obtenir une discrimination meilleure que 90 % sur des voyelles soutenues.

Les Figures I et II montrent comment se projettent les points utilisés sur les plans formés par les axes retenus, et comment se projettent les centres de gravité des classes sur ces mêmes plans. Un autre programme permet, avec ces résultats, de classifier une voyelle soutenue prononcée par un locuteur quelconque, masculin. Les résultats sont aussi de l'ordre de 90 % de bonne reconnaissance, même pour des locuteurs n'ayant pas fait parti du nuage de départ.

Une autre manière d'utiliser ces résultats est de considérer l'axe retenu comme étant la combinaison linéaire qui prend les valeurs les plus constantes pour une même voyelle en prenant les valeurs les plus distinctes pour des voyelles différentes. L'étude approfondie des variations de ce paramètre peut être alors une approche intéressante de la reconnaissance par éléments phonétiques.

IV - CONCLUSION .-

L'analyse discriminante linéaire peut être utilisée dans de nombreux autres domaines de reconnaissance de parole, soit qu'il s'agisse d'optimisation de paramètres, soit qu'il s'agisse de reconnaissance proprement dit dans un petit nombre de classes distinctes, dans la mesure où elle condense, en quelque sorte, l'information d'appartenance à une classe en éliminant au mieux l'information propre à chaque individu.

-o-

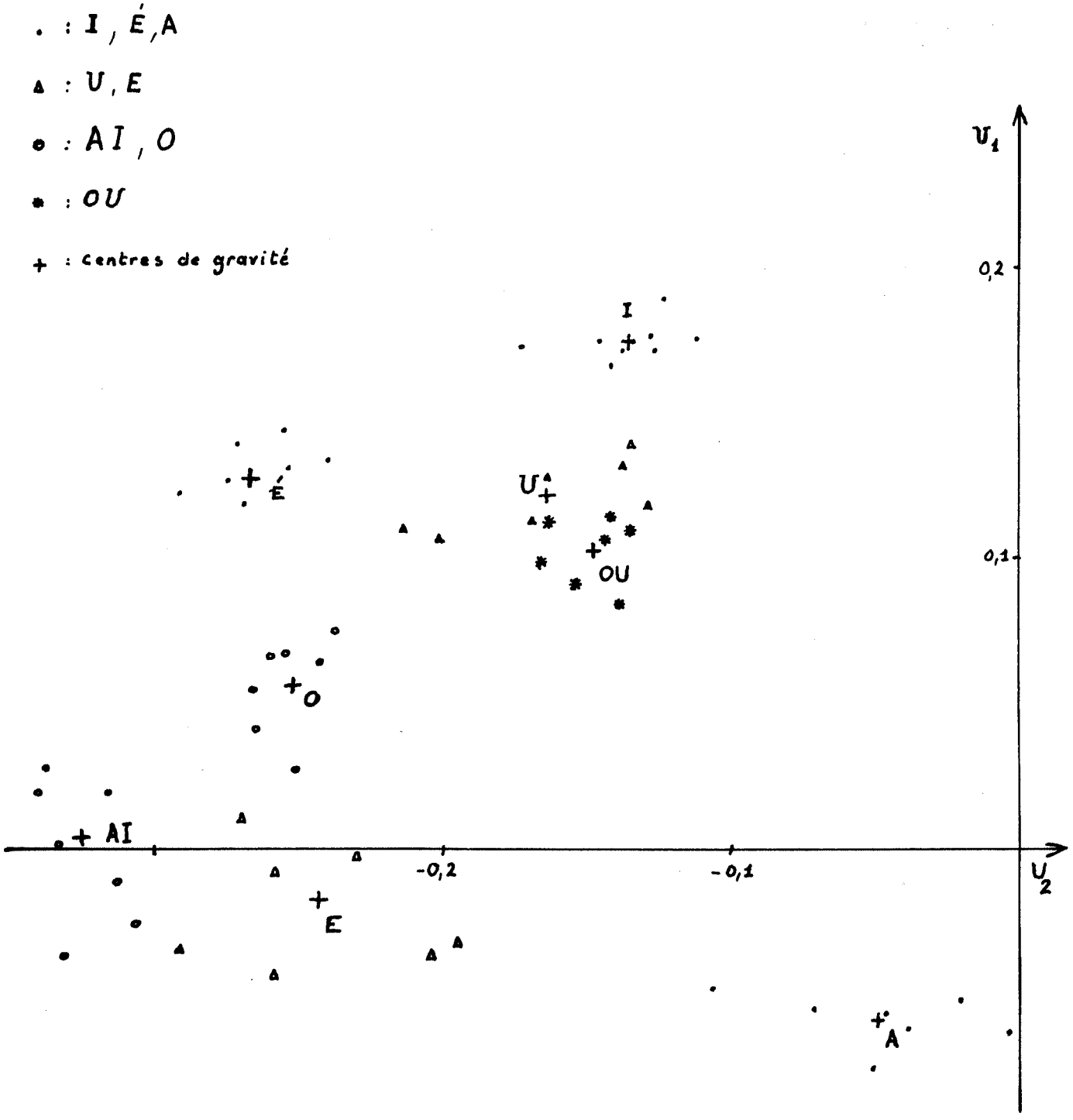


Fig 1 - Projections des vecteurs : spectres des voyelles dans le plan formé par les deux premiers vecteurs discriminants, ces vecteurs étant pris comme repère de ce plan.

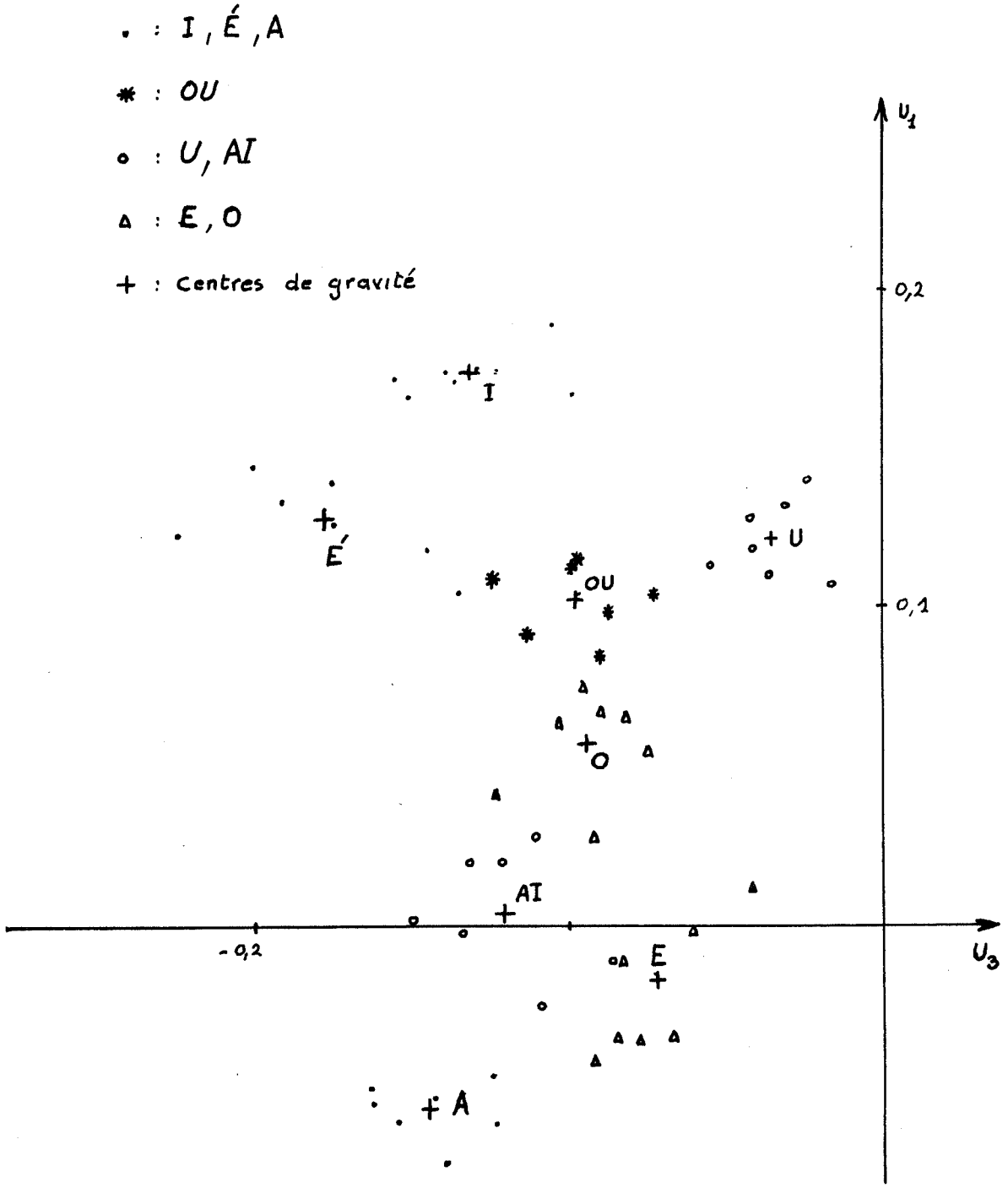


Fig 2 - Projections des vecteurs spectres des voyelles dans le plan formé par le premier et le troisième vecteurs discriminants, ces vecteurs étant pris comme repères du plan

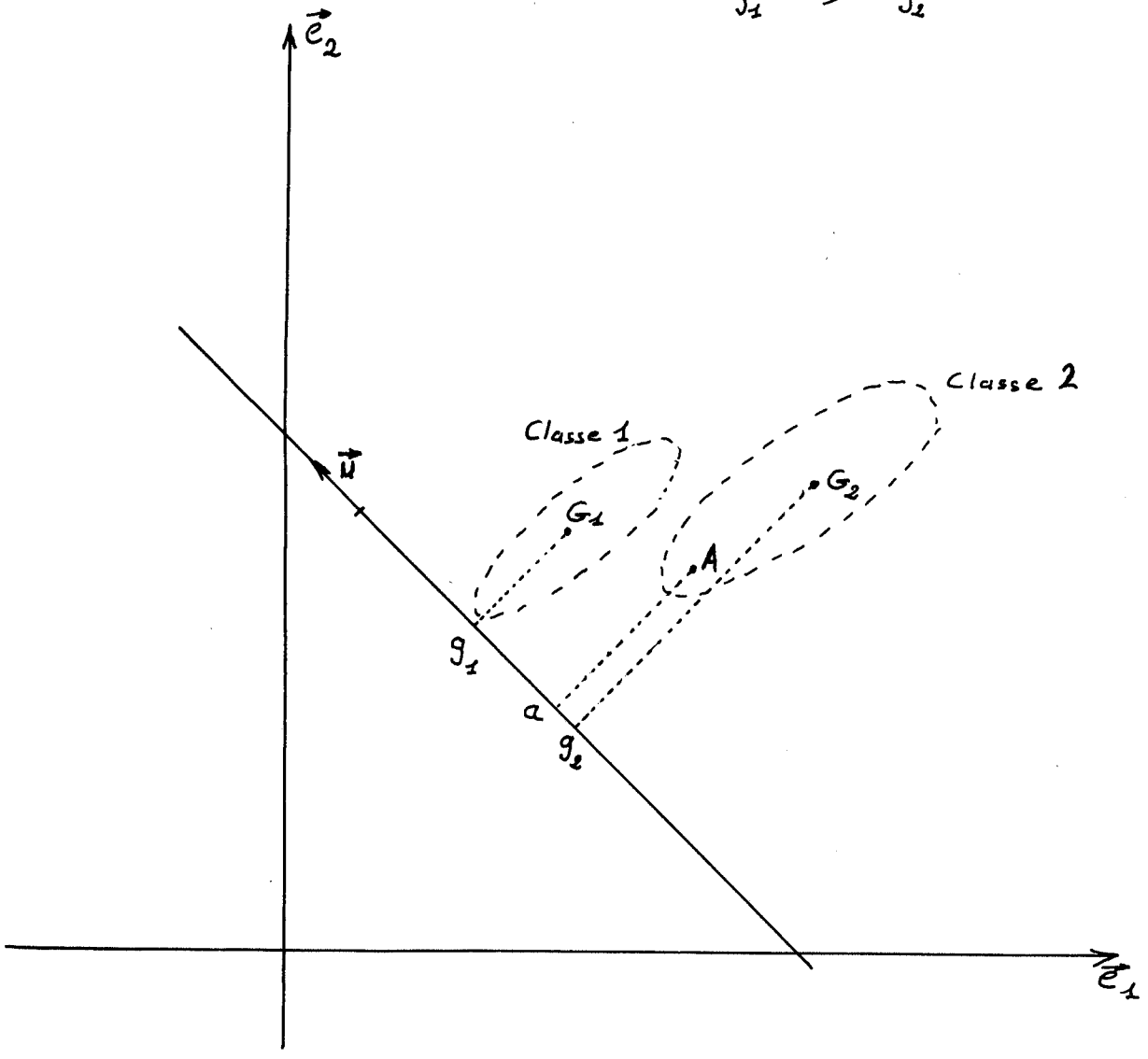
Figure III

A $\in$ Classe II

et : $\overline{G_1A} < \overline{G_2A}$ mais

en projection sur $\vec{u}$

$\overline{g_1a} > \overline{g_2a}$



RECONNAISSANCE GLOBALE DE LA PAROLE, EN TEMPS REEL, PAR CALCUL D'UN INDICE DE SIMILARITE FLOU

J. BREMONT et M. LAMOTTE
Laboratoire d'Electricité et d'Automatique
Université de NANCY I

Résumé

Nous reconnaissons, sur des données binaires, les mots d'un vocabulaire limité. Celui-ci est d'abord stocké par compression fréquentielle et temporelle en tenant compte de considérations "floues". Un calcul d'indice de similarité permet en phase de reconnaissance d'attribuer un score à chacun des mots du dictionnaire par comparaison au mot prononcé. Le meilleur score détermine le mot reconnu.

REAL-TIME SPEECH RECOGNITION WITH FUZZY SIMILARITY RATIO

Summary

A restricted number of words described by binary data are recognised. First each word of a learning sequence is compressed in time and frequency by using some fuzzy relations and is memorised. During recognition process the best similarity ratio between the uttered word and each word in reference gives the recognised word.

RECONNAISSANCE GLOBALE DE LA PAROLE EN TEMPS REEL PAR CALCUL D'UN INDICE DE SIMILARITE FLOUE

Nous reconnaissons les mots d'un dictionnaire limité. La parole est analysée puis codée par notre analyseur spectral [D1] qui délivre un échantillon de 24 bits toutes les 10 ms. Les mots stockés sont caractérisés par certains critères appliqués à des segments de parole voisins du phonème.

L'idée fondamentale du processus de reconnaissance repose sur le fait que, ne pouvant à priori prévoir certains caractères phonémiques ou acoustiques des mots en les particularisant trop, nous essayons au contraire de les envisager tous à la fois de façon la plus complexe possible sans les distinguer entre eux. Nous dirons que nous voyons les différentes zones du mot de façon "floue".

Cette façon de penser permet en conséquence de résoudre le problème de compression en temps et en fréquence des mots et à postériori également le problème de stockage des mots du dictionnaire sous forme réduite puisque nous pouvons alors envisager d'entrer un vocabulaire substantiel dans la mémoire d'un petit calculateur.

L'algorithme présenté procède en deux phases successives d'apprentissage et de reconnaissance.

Dans la première, nous construisons le dictionnaire en affectant au mot représenté en temps et en fréquence des degrés d'appartenance des éléments d'une partition de ce mot que nous appelons "pavés".

Dans la deuxième phase, l'algorithme, grâce au calcul d'un indice de similarité, compare les pavés ordonnés du mot prononcé par une excursion séquentielle au travers de chacun des mots du dictionnaire.

Le score le plus élevé sélectionne le mot reconnu.

A - Les données de l'analyseur-codeur.

L'expérience montre que deux occurences du même mot "Raoul" par exemple, obtenues sous forme binaire par notre analyseur [D1], se présentent différemment. Si, utilisant la première occurence comme référence nous recherchons à reconnaître la deuxième, nous constatons qu'il est nécessaire de s'affranchir des variations temporelles car les écarts peuvent atteindre de façon raisonnable 50 %.

Par ailleurs, si l'on regarde la zone correspondant au phonème /r/, on s'aperçoit que fréquentiellement et temporellement ces zones sont d'aspect différent mais doivent cependant apporter une contribution à la similarité des deux occurences.

Ces difficultés sont contournées de la manière suivante :

B - Brefs propos sur les ensembles flous.

En théorie de l'information, en linguistique, dans les problèmes d'apprentissage, en classification automatique, en reconnaissance de formes, etc..., les conditions du problème ou les objets ne sont pas toujours définis exactement mais il est cependant possible d'obtenir une représentation de ces données.

Par exemple, on peut parler dans un ensemble de nombres, de ceux qui sont "beaucoup plus grand" qu'un certain nombre N fixé ou bien des nombres "approximativement" égaux à N. Les termes soulignés sont des appréciations floues que nous pouvons formaliser par la théorie des sous-ensembles flous [K1], [Z1] dont nous ne précisons ici que la définition de base.

Soit l'ensemble $X = \{ x \}$. Le sous-ensemble flou S est donné par μ_S qui associe à chaque $x \in X$ un nombre réel de l'intervalle $[0, 1]$.

$\mu_S(x)$ est le degré d'appartenance de x à l'ensemble flou S.

Le choix de $\mu_S(x)$ est subjectif et n'est basé que sur des informations valables uniquement dans chaque cas particulier.

C'est une généralisation de la théorie des ensembles ordinaires dans laquelle nous introduisons pour $\mu_S(x)$ seulement les valeurs 1 et 0.

C - Considérations "floues" sur les formes acoustiques.

Nous admettons que les occurences à comparer d'un mot sont constituées de segments analogues voisins du phonème, placés relativement à la longueur du mot " à peu près " au même endroit si le locuteur parle normalement. "à peu près" est une hypothèse floue.

Nous appelons "zone temporelle" un segment de mot limité en général à des transitions entre phonèmes. Une "position relative" calculée par rapport à la longueur du mot marque la fin d'une zone temporelle définie.

Nous cherchons alors à caractériser, dans une zone temporelle, l'aspect fréquentiel d'un segment de mot. Nous envisageons des groupements de "1" pour plusieurs ensembles de filtres. Nous avons retenu par exemple la partition de 8 filtres pour les hautes fréquences, 8 filtres pour les moyennes fréquences et 8 filtres pour les basses fréquences dans l'ensemble des filtres.

Chaque zone est de plus caractérisée par un ensemble de paramètres définis par rapport à certains critères.

La forme globale d'un mot se présente comme une mosaïque de petits "pavés" élémentaires définis par l'intersection des zones temporelles et des bandes de fréquences.

Nous appelons "critère" le comptage des "1" dans un tel rectangle.

L'importance du groupement de "1" présents dans le pavé d'une zone temporelle est évaluée par la mesure de l'appartenance au critère amenant ainsi une compression de l'information pour la zone temporelle. Cette dernière est alors constituée par la combinaison des degrés d'appartenance aux critères de la zone temporelle.

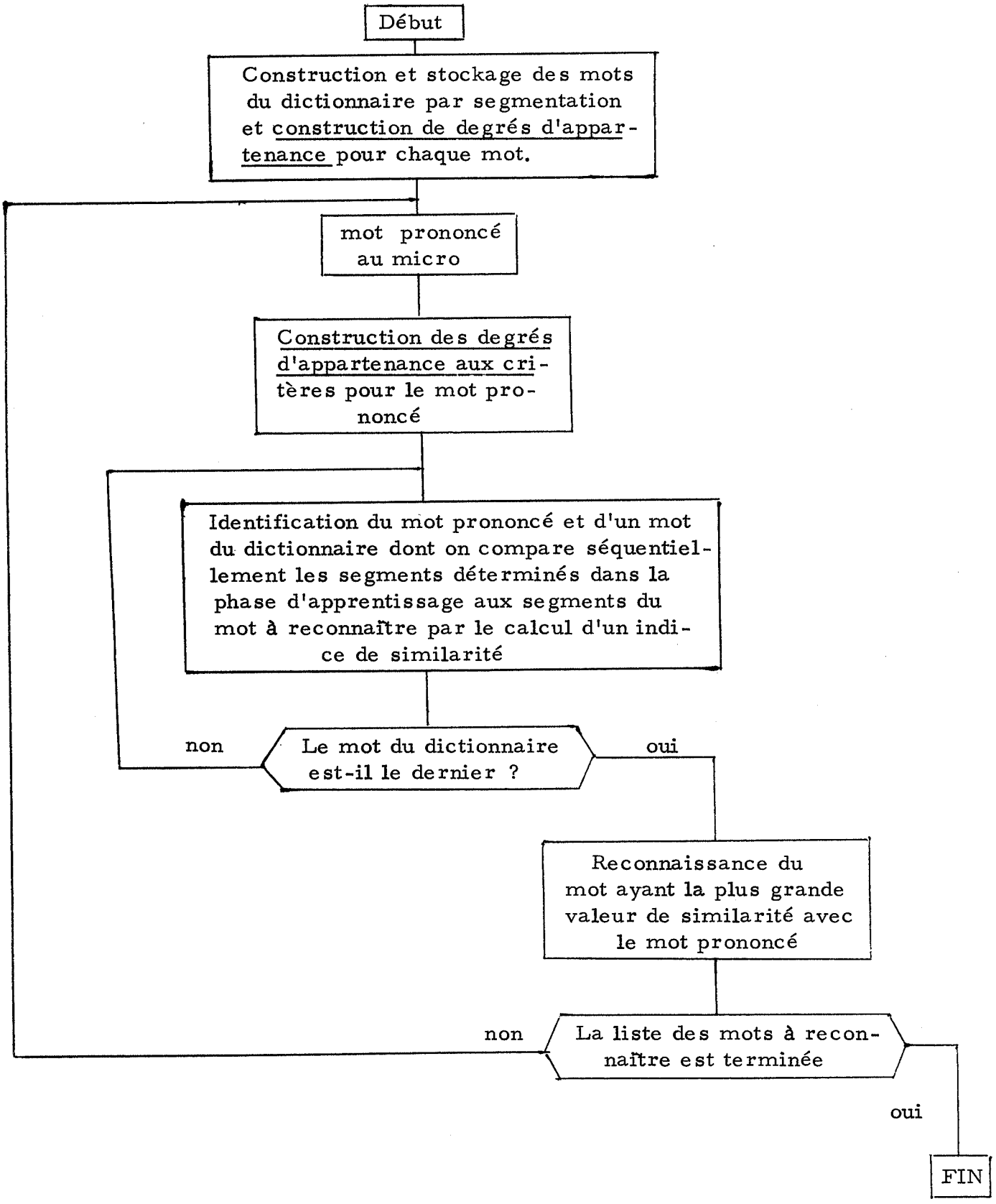
Un mot est défini par un ensemble de pavés traduisant, sous forme condensée, l'aspect acoustique de ce mot. Or, chaque mot du "langage" est défini par une chaîne ordonnée de segments ; il est donc normal d'organiser les pavés en apportant cette nouvelle source d'information dans le processus de reconnaissance, dès lors que la liste des mots à reconnaître est fixée.

Nous pouvons alors envisager le processus de reconnaissance.

D - Processus de reconnaissance.

Il est décrit dans l'organigramme suivant :

D1 : Organigramme simplifié :



D2 : Calcul de la position relative et des degrés d'appartenance dans la phase d'apprentissage et celle de reconnaissance :

En phase d'apprentissage on ne mémorisera pour un mot que la position relative de fin de zone temporelle. Cette position relative s'exprime par :

$$N R_i = \frac{N_i \times RF}{ECH} \quad \text{pour } i = \{1, \ldots, NZ\}$$

où RF est une longueur standard
N_i la position absolue de la fin de la zone i du mot
ECH le nombre d'échantillons du mot
NZ le nombre de zones du mot.

Le degré d'appartenance du nombre de points NP_i présents dans un critère s'écrit :

$$\mu^i_{jm}(NP_i) = \frac{NP_i}{(N_i - N_{i-1}) \times VM}$$

où VM est le nombre de filtres dans une zone fréquentielle.

C'est le degré d'appartenance des points NP_i au pavé défini par la zone temporelle i et la zone fréquentielle j pour le mot m.

D3 : Phase d'apprentissage :

Cette phase nous permet de construire le dictionnaire de tous les mots prononcés successivement au micro donc de ranger les mots de référence. Un mot est segmenté automatiquement [L1]. La forme et les points de segmentation sont visualisés. Si le mot est jugé convenable, l'algorithme précédent calcule la position relative correspondant aux points de segmentation et calcule le degré d'appartenance de chaque pavé. Ces résultats sont stockés en mémoire ainsi que le libellé du mot.

Nous pouvons alors considérer la phase de reconnaissance.

E - Phase de reconnaissance proprement dite.

Le locuteur prononce successivement les mêmes mots que ceux qui ont été stockés sous forme réduite dans le dictionnaire lors de la phase précédente d'apprentissage.

Le calculateur cumule pour chaque critère les "1" présents dans le dernier échantillon acquis et les précédents, et calcule les

positions relatives de chaque échantillon par rapport à la référence RF avant de passer à l'algorithme de reconnaissance.

Le mot prononcé est alors comparé aux seules positions relatives de chacun des mots du dictionnaire suivant l'indice de similarité suivant :

$$S(z_{ik}, z_{ik}) = 1 - \frac{\sum_{j=1}^{CRIT} \left| \mu^{i}_{jk}(NP_i) - \mu^{i}_{j\ell}(NP_i) \right|}{CRIT}$$

c'est-à-dire pour une zone temporelle i avec un nombre "CRIT" de critères. Si k désigne le mot du dictionnaire et ℓ le mot prononcé, z_k et z_ℓ sont alors les zones temporelles se correspondant respectivement.

F - Résultats obtenus [B1].

Trois listes d'occurence de 19 mêmes mots ont été perforées sur bande pour être répétitifs et ont permis de constater des pourcentages de reconnaissance dépendant du nombre de critères choisis. Ils sont donnés dans le tableau suivant :

nombre de critères	pourcentage de reconnaissance
1	79
2	95
3	100
4	100
6	100
8	100
12	100

De plus, la discrimination entre le mot prononcé reconnu et les autres mots du dictionnaire augmente avec le nombre de critères.

Le choix de quatre critères, suffisant pour être efficace, nous a permis d'écrire le programme sur "T 2000" (Télémécanique Electrique), sous une forme peu encombrante. Il satisfait le stockage et la reconnaissance d'à peu près 200 mots (avec 1820 mots

de 19 bits pour le programme proprement dit et 5500 mots de 19 bits pour le volume des tableaux tels que le libellé et le dictionnaire).

Conclusions et perspectives.

Dans l'objectif de reconnaissance des mots par comparaison d'un mot prononcé par rapport à ceux de référence, la méthode résoud de façon très intéressante le problème de la compression temporelle et fréquentielle des mots. De plus, ce qui n'est pas un mince avantage sur d'autres méthodes efficaces existantes, elle permetde stocker un dictionnaire important dans la mémoire d'un petit calculateur. De tels résultats laissent envisager la suite de ce travail sur la reconnaissance de la parole continue.

Références.

/B1/ BREMONT Jacques "Apport de "considérations floues" à la paramétrisation et à la reconnaissance automatique de la parole" . Publication interne au L. E. A. (NANCY I) (11. 3. 1974).

/D1/ DUGRAVOT M. -J. "Prétraitement numérique du signal vocal dans le domaine spectral" Thèse de Docteur en spécialité Automatique - Université de Nancy I (25 octobre 1969).

/K1/ KAUFMANN A. "Introduction à la théorie des sous-ensembles flous" Tome 1 - Masson et Cie (1973).

/Z1/ ZADEH L. -A. "Fuzzy Sets" , Information and Control V8, P. 338-353 (1965).

/L1/ LAMOTTE M. "Segmentation" Publication interne au L. E. A. (NANCY I) (11. 3. 1974).

- RECONNAISSANCE DE LA PAROLE - PAR ALGORITHME D'APPRENTISSAGE D'OPERATEURS

C.ROCHE, F.CHATEAUNEUF
LCA - Fort de Montrouge 94110 ARCUEIL

RESUME : Le but de nos expériences est de reconnaître 200 mots prononcés par différents locuteurs et utilisant un vocodeur à 18 canaux peu performant. L'étude porte principalement sur l'utilisation de la redondance à tous les niveaux de reconnaissance. L'apprentissage d'opérateur est utilisé pour la reconnaissance des phonèmes. La segmentation a pour but de positionner les opérateurs de reconnaissance de phonèmes aux instants où ils ont le taux d'erreur le plus faible. La reconnaissance des mots utilise une méthode de recherche avec backtrack et choix des questions les meilleures à poser.

ABSTRACT : Speech recognition with learning of recognition operators. Thegoal of our experiments is to recognize 200 words pronounced by different speakers using a poor quality 18 channel vocoder. The study is principally about the use of the redundancy at each recognition level. Learning of recognition operators is used for the phonemes. A segmentation algorithm is positionning the phoneme recognition operator so that they have the lowest error rate. The recognition of the words use a backtrack-forward looking method for the choice of the best questions to ask.

I/- INTRODUCTION

Nous nous proposons l'étude d'un Système de Reconnaissance de mots, issus d'un vocabulaire limité (500 mots), prononcés par un ensemble fini de locuteur (10 au maximum), en temps réel. Le programme de reconnaissance basé sur la notion d'apprentissage, utilise des méthodes statistiques empruntées plus particulièrement à la théorie de l'information (information utile). Le projet est divisé en trois parties :

1) - la reconnaissance des phonèmes,

2) - la segmentation du mot en phonèmes,

3) - la reconnaissance du mot,

la première partie étant en cours d'achèvement.

II/- RECONNAISSANCE DES PHONEMES

II.1 - Acquisition des données - Restitution sur vocoder

Elle se fait en temps réel sur un vocoder aux performances moyennes voire médiocres, composé d'un extracteur de pitch et d'une série de 18 filtres ($\Delta f / f = CTE$) codés en modulation delta (pas = 25 ms). Un programme permet d'extraire à partir des séquences codées de l'acquisition les trames de 18 énergies et le pitch formant le sonogramme.

Inversement nous pouvons resynthétiser la parole à partir des sonogrammes obtenus en première analyse.

II.2 - Etablissement des données d'apprentissage

Nous avons fait l'acquisition sur vocoder d'environ $\frac{1}{2}$ heure de parole par 6 locuteurs. Les conditions d'enregistrement ont varié sensiblement selon le locuteur : élocution rapide, normale, milieu calme, bruité (salle d'ordinateur), voix d'intensité variable.

La phase d'apprentissage consiste à découper le sonogramme en segments matérialisant des phonèmes et à identifier chaque segment manuellement. Le nombre élevé de trames

(7000) rend difficile le pointage sur un listing du sonogramme, d'autant plus difficile qu'il est préférable que le travail soit fait par le même opérateur.

Pour faciliter la tache nous faisons défiler le sonogramme sur console de visualisation et nous pointons au light-pen le début de chaque phonème reconnu par l'opérateur, puis nous tapons sur le clavier le nom du phonème et le nombre de trames occupées. L'automatisation d'une part importante du travail de l'opérateur permet d'accélérer notablement la phase d'apprentissage.

II.3 - Principe de la méthode d'apprentissage

On généralise la notion d'opérateur qui est représenté par les noms de deux variables d'entrée, le nom de la variable de sortie et un tableau à double entrée. Le résultat de l'opération sur deux variables est la valeur de la case du tableau indexé sur les lignes et les colonnes par les valeurs des deux variables d'entrée. L'utilisation en chaîne de ces tableaux permet de traiter un ensemble de variables représentatives d'une forme et finalement d'identifier chaque réponse du dernier opérateur à un phonème différent. Un programme général permet de créer cette succession d'opérateurs et leurs tables. Cette génération repose sur des critères statistiques, sur des notions empruntées à la théorie de l'information (information utile, entropie, conditionnelle). Voir [1] et [2].

II.4 - Application de la méthode dans un cas général

Dans une première approche, nous avons tenté d'évaluer les difficultés et les problèmes posés par la reconnaissance des phonèmes et testé le programme de génération automatique d'opérateurs.

Un certain nombre d'expériences ont été faites en prenant pour paramètres d'entrée de la structure de reconnaissance les valeurs mêmes de sortie du vocodeur.

Nous avons noté une corrélation importante entre les canaux consécutifs, un gain d'information faible à chaque étape de regroupement de variables et finalement des taux de reconnaissance différents sur l'apprentissage et sur l'ensemble de test. Malgré un choix de variables peu interessantes, nous notons une confusion relativement faible entre les classes classiquement définies (voyelles, fricatives....). Le taux d'erreur varie de 30% pour les fricatives à 80% pour les plosives (50% sur l'ensemble des phonèmes).

II.5 - Les phonèmes soutenus

Un examen rapide du sonogramme fait apparaitre en première analyse deux classes de sons :

- les sons quasi stationnaires ou soutenus caractérisés par des formants stables
- les sons transitoires, très peu marqués parfois même inexistants sur le sonogramme.

Actuellement l'étude des sons soutenus englobant les voyelles et les fricatives, est terminée.

a) - Les voyelles (au nombre de 14)

Dans l'étude des sons soutenus, définis par la position de leur formants, considérant la faible évolution fréquentielle de ces formants, nous prenons une décision par trame.

Nous avons débuté l'analyse des voyelles par une tentative de reconnaissance basée sur les 18 énergies de la trame comme variables d'entrée de la structure de reconnaissance.

Sur l'apprentissage :	50%	de bonne décision
	70%	si l'on tolère deux éventualités pour chaque réponse.
Sur le test	35 %	
(locuteur différent	60 %	
(mots différents)		

Dans une autre série d'expériences nous nous sommes attachés finalement à une description formantique de la structure des voyelles. Après un lissage grossier du sonogramme, rendu nécessaire par les incertitudes introduites par le fonctionnement du capteur, nous associons 1 formant à tout maximum du spectre instantané. Nous localisons également les minima locaux et effectuons une reconstitution du spectre instantané par interpolation linéaire sur les extrêma . Une restitution parfaitement intelligible sur vocoder nous assure de la conservation de l'information utile. Nous décrivons alors chaque formant par un ensemble de variables, (8 par formant) et un traitement adapté est effectué suivant le nombre de formants. La mauvaise résolution fréquentielle du vocoder apparait sur les histogrammes des fréquences centrales très confus. Le programme de génération d'opérateurs donne des gains importants d'information lors du regroupement de 2 variables et donne toujours un certain décalage entre les pourcentages de bonne décision sur l'apprentissage et sur le test : 55% et 35%.

Le premier programme de génération d'opérateurs ne peut traiter qu'un nombre limité de variables et les fait toutes intervenir pour la décision, quelque soit leur qualité.

Un deuxième programme, inspiré du précédent nous permet de traiter un nombre élevé de variables, et de ne choisir que les plus informantes. Nous décidons alors d'appliquer des opérateurs simples aux variables de base décrivant les formants : +, -, x, / , comparaison, barycentre. Nous engendrons ainsi de nouvelles variables et constatons que les meilleures apportent des informations identiques pour chaque locuteur. Toutefois, on obtient le même pourcentage de bonne décision que pour le premier programme. On obtient donc en moyenne sur toutes ces expériences un pourcentage de 50% par trame, qui se bien entendu majoré sensiblement lors de la décision sur plusieurs trames, au niveau du phonème.

b) - Les fricatives (au nombre de 6)

Il convient d'ajouter aux variables précédentes, le pitch qui est assez informant pour les fricatives. Les histogrammes de fréquences centrales d'énergies de formants sont plus différenciés que pour les voyelles et expliquent le pourcentage satisfaisant de décision. Le traitement a été identique à celui appliqué aux voyelles.

Variables d'entrée : les 18 valeurs de sortie du vocodeur	- 75% de bonne décision (mais différence de pourcentage avec le test)
Variables d'entrée les mêmes que pour les voyelles	- 90% pour les fricatives présentant un formant
	- 70% pour les autres
Performances sur le test (locuteur différent)	- 40%

c) - Les plosives

La reconnaissance des plosives est en cours d'étude.

III/- LA SEGMENTATION

Le principe de la méthode de reconnaissance de phonèmes est d'activer les opérateurs de reconnaissance à chaque digitalisation du spectre. Ces opérateurs travaillent sur la trame de digitalisation et le passé immédiat de cette trame, soit 3 trames ou 75 ms.

Les opérateurs de segmentation que nous cherchons sont ceux qui détectent les positions optimales des phonèmes ou qui donnent des taux d'erreur minimaux pour les performances des opérateurs de reconnaissance de phonème.

Dans un premier temps la recherche de ces opérateurs de segmentation se fait par l'informaticien guidé par son expérience du phénomène acoustique.

Des fonctions f1, f2, f3, f4 détectent des notions intuitives correspondant à des formes d'une complexité moindre que la notion de phonèmes.

f1 et f2 représentent l'énergie; f3 la stabilité; f4 l'énergie dans les hautes fréquences. Le but étant de formaliser les points intuitifs et d'expérience suivants :

Une plosive présente un maximum d'instabilité avec un accroissement de l'énergie. Une voyelle, un maximum d'énergie et de stabilité. Une sifflante un maximum d'énergie dans les hautes fréquences et un maximum de stabilité.

L'informaticien a donc dans un deuxième temps cherché les meilleures fonctions f1, f2, f3, f4, par un tatonnement systématique utilisant comme critère l'optimisation des performances des opérateurs de reconnaissance de phonèmes.

Ces fonctions ne peuvent avoir rien de général : elles dépendent de manière très étroite du capteur et de ses valeurs de sortie.

Un certain nombre de filtres ont été essayés, les fonctions fi sont des fonctions simples et arithmétiques des résultats de filtrage. Pour un gain de performance il a fallu considérer deux fonctions f4 de stabilité, l'une pour la détection des voyelles, l'autre pour les consonnes.

Les performances obtenues sont approximativement de 70% de détection de voyelles, 50% de sifflantes et 80% de plosives. Ces performances peuvent encore être considérablement améliorées.

Cette recherche par tatonnement peut être automatisée. Ce système d'apprentissage est en cours d'étude et donnera sans doute de meilleures performances au programme de segmentation, le nombre d'essais pouvant être effectués étant supérieur de plusieurs ordres aux nombres d'essais effectués à la main pour un moindre coût. L'apprentissage est possible ici : le critère a été défini, ainsi que l'espace des fonctions fi.

IV/- LE PARTAGE EN CLASSES DE RYTHME VOISIN

Le rythme du mot est donc défini par l'ensemble des fonctions fi. Nous nous proposons d'étudier l'influence du rythme sur la reconnaissance des mots et d'en tenir compte pour améliorer les performances du système de reconnaissance des mots.

Le programme de reconnaissance des mots (§ V) travaille sur les résultats des opérateurs de reconnaissance de phonèmes positionnés par les fonctions de segmentation. Ce programme ne peut travailler avec un temps acceptable que sur un nombre de mots faibles (20 à 40). Si on veut reconnaître 200 mots, il faut donc diviser ces mots en classes plus petites par un premier étage de reconnaissance et ensuite affiner la recherche par l'algorithme décrit au paragraphe V.

Cette classification s'effectue par la valeur des paramètres de rythme.

Des fonctions fi, on déduit pour un mot prononcé les valeurs de paramètres décrivant au mieux ces fonctions; c'est-à-dire tels qu'il soit possible de reconstituer les allures des fi à partir des valeurs de ces paramètres. Il faut par ailleurs que ces valeurs soient invariantes par rapport aux déformations permises en amplitude et en temps : voir [7]

On définit alors une distance de ressemblance entre mot prononcés par la valeur de :

$$d(i,i') = \sum_{j} (p_i^j - p_{i'}^j)^2$$

après normalisation des $p_i j$ pour rendre leur variance statistique constante. Sur un ensemble d'apprentissage d'une centaine de mots on utilise un algorithme de classification a priori : la méthode employée est celle des Nuées Dynamiques de E. Diday.

Les premiers résultats montrent une dispersion de tous les mots et des classes peu nettement séparées.

Cette répartition uniformément répartie des mots définis par leur rythme est à rapprocher d'une tendance naturelle qu'on constate en reconnaissance des formes à la maximisation de l'entropie concernant les formes interessantes : voir [1]

Le partage en classes s'effectue de manière correcte en considérant les classes avec recouvrement : voir [7] . Les classes de rythme voisin ont alors de l'ordre d'une quinzaine de mots, ce qui était le but recherché.

V/- LA RECONNAISSANCE DES MOTS

La décision de l'appartenance de telle ou telle classe de rythme voisin, se fait par le minimum de la moyenne des distances aux noyaux des classes (formes typiques représentatives de chaque classes et obtenues automatiquement par l'algorithme des nuées dynamiques).

A l'intérieur de chaque classe de rythme voisin, la reconnaissance du mot s'effectue par l'interrogation successive des résultats des opérateurs de reconnaissance de phonème. La méthode utilisée est à rapprocher de celle décrite en [5]

La méthode décrite en [5] permet d'évaluer à priori la meilleure question à poser séquentiellement pour reconnaître la forme. L'évaluation se fait par une mesure de l'information utile apportée par chaque question pour reconnaître les formes données : voir [1] . Cette fonction est mesurée sur un ensemble d'apprentissage. A chaque réponse les probabilités d'avoir telle ou

telle forme se modifient (formule de Bayes) et lorsque la probabilité d'erreur est inférieure à une valeur donnée, la meilleure décision est prise. La méthode utilisée ici tend à être plus sophistiquée : l'évaluation des meilleures questions se fait non plus coup par coup mais comme en théorie des jeux en analysant le questionnaire plusieurs questions d'avance. La méthode est due à J. Slagle [9] , la reconnaissance est assimilée à un jeu contre la nature, dans lequel le but du joueur est de reconnaître les formes, donc arriver au résultat dans le minimum de coups, mais dans lequel la nature n'a pas une réponse la plus défavorable pour nous mais aléatoire. Les procédures classiques en théorie des jeux du minimax ou alpha-beta sont remplacées ici par le maximoyen et gamma : voir [9] et [8]

L'algorithme n'a été testé que sur des données simulées : suite de phonèmes présentant un pourcentage d'erreur considérable. Les matrices de confusion des opérateurs de reconnaissance de phonèmes ont elles aussi été simulées, à partir de résultats antérieurs correspondant à des taux de bonne reconnaissance de phonèmes de l'ordre de 40%.

Dans ce cas et lorsque le mot à reconnaître était à discriminer parmi une vingtaine de mots, les performances de l'algorithme ont été comparables aux performances de l'oreille humaine : 80% de bonne reconnaissance.

VI/- CONCLUSION

Notre but n'est ni la compréhension du phénomène acoustique, vocal ou phonétique, ni une reconnaissance parfaite des mots mais : utilisant des capteurs et des systèmes de traitement approximatifs, et peu couteux, de les organiser en les optimisant pour arriver à un système utilisable travaillant sur 200 mots et plusieurs locuteurs.

A cette occasion nous avons essayé de mettre en oeuvre des idées propres à la reconnaissance des formes, que nous utilisons par ailleurs en reconnaissance de formes visuelles et traitement d'image : l'algorithme de segmentation par exemple se retrouve en traitement d'image sous le nom "détection automatique de points caractéristiques sur photographies"....

Peut être alors ces méthodes de reconnaissance des formes pourront elles servir d'outils à la compréhension du phénomène vocal, à la définition des sons élémentaires, à la détermination des modèles de perception.... ou simplement être utilisées dans la conception de systèmes de reconnaissance plus performants avec des capteurs plus précis.

REFERENCES :

[1] C.ROCHE : Information Utile en Reconnaissance des Formes et en Compression de données. Application à la Generation Automatique de Systèmes de Reconnaissance Optique et Acoustique. Thèse Dec.72.

[2] C.ROCHE : Application of Multilevel clustering to the the Automatic Genration of Recognition Operators. Alink between Feature Extraction and Classification. Présenté au IJCPR Copenhagen - Août 1974.

[3] JC.SIMON: "Application of Questionnaire Theory to Pattern C.ROCHE Recognition" IJCAI Sept.71.

[4] JC.SIMON, C.ROCHE, G.SABAH : "On automatic Generation of Pattern Recognition Operators" International Conference on Cybernetics. Oct. 72.

[5] C.ROCHE : "Idées Générales sur la Reconnaissance des Formes appliquées à la reconnaissance de la Parole" journées du GALF. Aix-en-Provence Mars 71 - Automatisme Mars 72.

[6] C.ROCHE : "Generation automatique d'opérateurs de segmentation de la parole" journées du Galf - Lannion Juin 72.

[7] LEROI : Rapport de Dea. Institut de Programmation Université Paris VI 1973.

[8] DUPOUY, YELLOZ : idem 1973.

[9] J. SLAGLE, R.LEE : "Application of Game Tree Searching Techniques to Sequential Pattern Recognition" Commun of the ACM - Février 1971.

Reconnaissance par mots basée sur les transitions phonétiques

M. MLOUKA - J.S. LIENARD
L.I.M.S.I. du C.N.R.S. - B.P. 30 - 91406 ORSAY

Résumé :

La reconnaissance de la parole est généralement abordée à partir de la notion de phonème. Nous présentons ici les premiers résultats d'une approche différente, dans laquelle seules les transitions sont prises en compte. Le mot à reconnaître est comparé aux mots-références selon un processus de comparaison dynamique. Le taux de reconnaissance obtenu par un locuteur unique est supérieur à 90% pour 100 mots et une seule passe d'apprentissage.

Word recognition based on phonetic transitions

Summary :

Speech recognition is generally handled with the concept of phoneme. Here are presented the first results of a different approach, in which only the transitions are considered. The word to be identified is compared to the stored reference words by means of an elastic matching process. The correct response percentage for a single speaker is above 90% with 100 stored words and one single training set.

Reconnaissance automatique de mots basée sur les transitions phonétiques (*)

M. MLOUKA - J.S. LIENARD

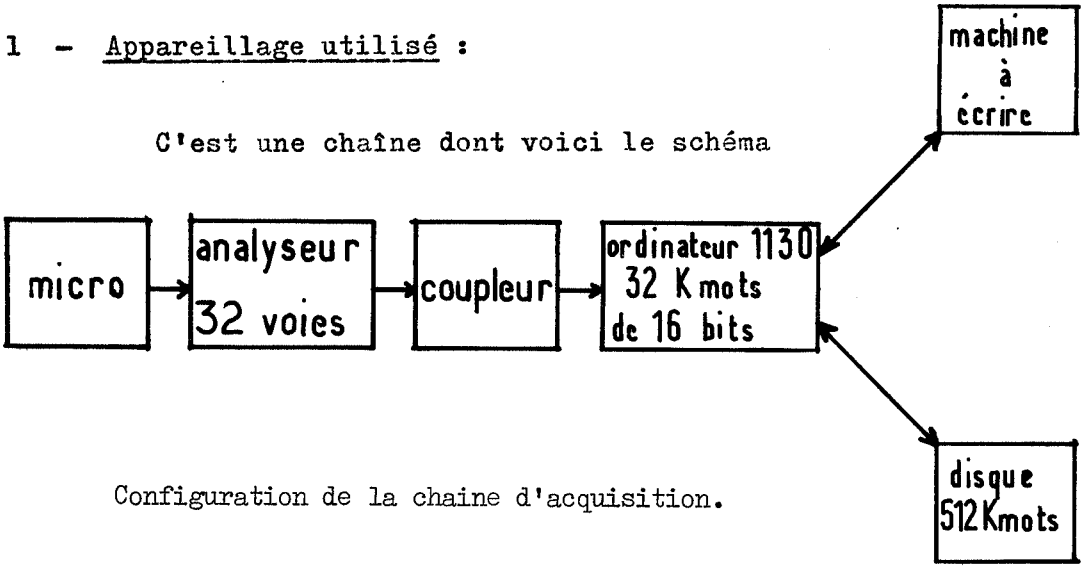
LIMSI - CNRS

L'expérimentation décrite ci-après a pour but de montrer que la reconnaissance de la parole peut être abordée avec succès à partir de la seule connaissance des états transitoires de la parole. La parole n'est plus décrite en termes de fréquence et amplitude de formants, mais en termes de variations spectro-temporelles. Les idées de base de cette étude sont analogues à celles que nous développons en synthèse (1).

Nous décrirons successivement l'appareillage d'analyse disponible au L.I.M.S.I., le principe de l'expérimentation (reconnaissance par mots), le codage des transitions et les résultats obtenus.

1 - Appareillage utilisé :

C'est une chaîne dont voici le schéma



Configuration de la chaine d'acquisition.

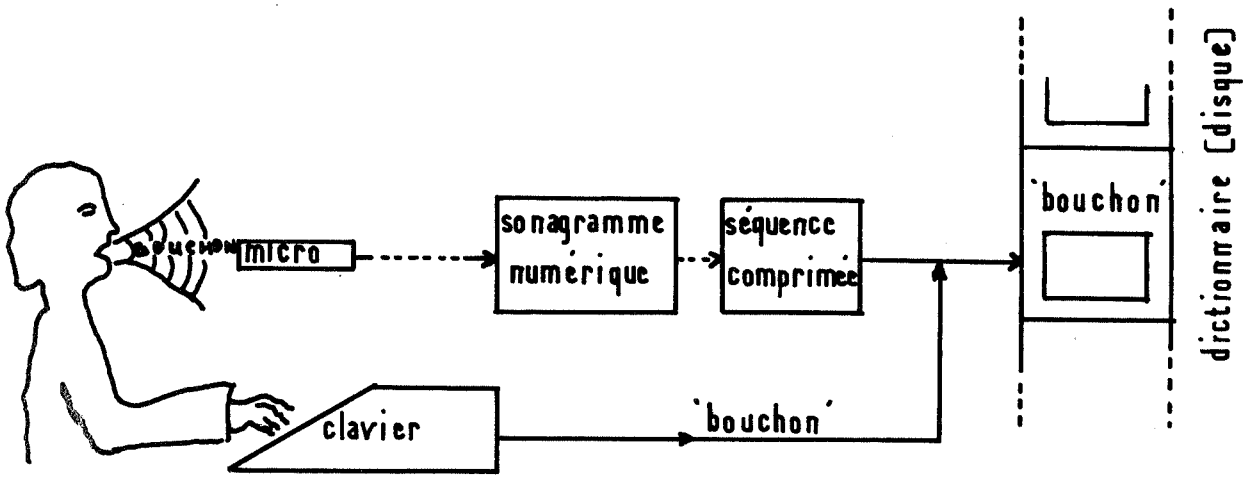
(*) : étude menée dans le cadre du contrat DRME 73/169

L'analyseur est un banc de 32 filtres analogiques. Dans son réglage actuel, il couvre la plage de fréquence 100 Hz - 7000 Hz. La répartition des fréquences centrale est logarithmique au dessus de 600 Hz, linéaire au dessous. La largeur de bande est pratiquement constante jusque vers 3000 Hz et croissante au delà. L'analyseur fournit toutes les 10 ms, 32 valeurs spectrales comprises entre 0 et 1023.

2 - Principe de l'expérimentation

a) Constitution du dictionnaire de référence

Un locuteur prononce un mot devant le microphone. Un programme de prétraitement permet d'éliminer les instants de silence qui précèdent et suivent la séquence utile. La suite de spectres numériques est ensuite transformée en une séquence comprimée selon un codage spécial, dont nous verrons le détail plus loin. L'opérateur fournit l'étiquette correspondante au programme qui la range sur disque avec la séquence comprimée.



Phase de constitution du dictionnaire: Un locuteur prononce un mot (ici"bouchon") devant le microphone , puis introduit au clavier la chaine alphabétique correspondante . Le système repertorie ces informations dans le dictionnaire .

b) Reconnaissance

Le mot à identifier est comprimé de la même manière que les mots-références, puis comparé à tous les éléments du dictionnaire. Le mot reconnu est celui qui donne lieu à la meilleure note de ressemblance. La recherche dans le dictionnaire n'est pas optimalisée, l'objet essentiel de l'expérimentation étant d'éprouver la validité du codage adopté.

3 - Compression et codage

Lors d'expérimentations antérieures (4) la séquence comprimée était constituée des spectres correspondant alternativement aux maxima et aux minima de stabilité spectrale.
Les idées développées en synthèse tendaient cependant à accorder plus de valeur perceptive aux transitions, c'est-à-dire aux instants de plus grande instabilité. C'est pourquoi nous avons comparé, toutes choses égales par ailleurs, la valeur des spectres stables d'une part, et celle des vecteurs représentant les transitions d'autre part, au regard de la reconnaissance par mots. Les vecteurs de transition étaient fournis par la simple différence des spectres entourant les instants de plus forte instabilité. Cette expérimentation nous a montré que les vecteurs de transition étaient plus significatifs que les spectres stables (2).

Nous avons donc décidé de baser notre système de reconnaissance uniquement sur les vecteurs de transition, en améliorant leur description. Le traitement actuel est le suivant :

a) Transformation de la suite de spectres numériques en suite de vecteurs de transition (vecteurs "Δ")

soit A le sonagramme numérique constitué des $a_{i,j}$ où i représente le temps et j la voie d'analyse.

soit Δ un vecteur de composantes :

$$\delta_1 , \delta_2 , \delta_3 , \ldots\ldots\ldots , \delta_{32}$$

pour un instant i et un filtre j

on a $$S1_j = a_{i+n,j}$$

et

$$S2_j = a_{i-n,j}$$

l'intervalle de temps 2n est choisi égal à 40 ms

on a alors $$\delta_j = \frac{S2_j - S1_j}{S2_j + S1_j + B_j}$$

où Bj est le bruit de fond moyen sur la voie j .
Ce bruit de fond est calculé pendant la période de silence relatif précédent et suivant la séquence utile. Il est bon de remarquer que le calcul des δ_j réalise une normalisation en niveau et tient compte du bruit de fond ambiant.

b) Segmentation et choix des instants importants.

On fait correspondre à cette séquence de vecteurs "delta" une fonction d'instabilité, dont la valeur est donnée à chaque instant par

$$\sum_{j=1}^{j=32} |\delta_{i,j}|$$

On forme la séquence comprimée en sélectionnant les vecteurs correspondant aux maxima successifs de cette fonction.
La compression est de l'ordre de 90%.

4 - <u>Détermination d'une note de ressemblance</u>

a) Note de ressemblance de deux vecteurs delta

soit $\Delta 1$ et $\Delta 2$ ces deux vecteurs.

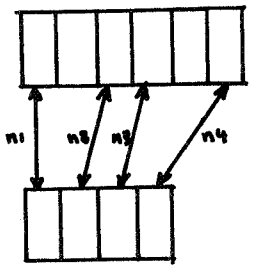
Nous avons choisi la distance conduisant aux calculs les plus simples ; la note de ressemblance est définie par :

$$\sum_{j=1}^{j=32} |\delta 1_j - \delta 2_j|$$

b) Note de ressemblance de deux séquences comprimées

Il s'agit d'une comparaison dont la progression temporelle est dynamique pour la séquence la plus longue. En effet on progresse par pas de temps réguliers dans la plus courte des deux séquences. A chaque vecteur de celle-ci on permet au système d'associer le vecteur le plus ressemblant parmi deux vecteurs successifs de la plus longue. C'est celui-ci qui va déterminer la progression temporelle pour le

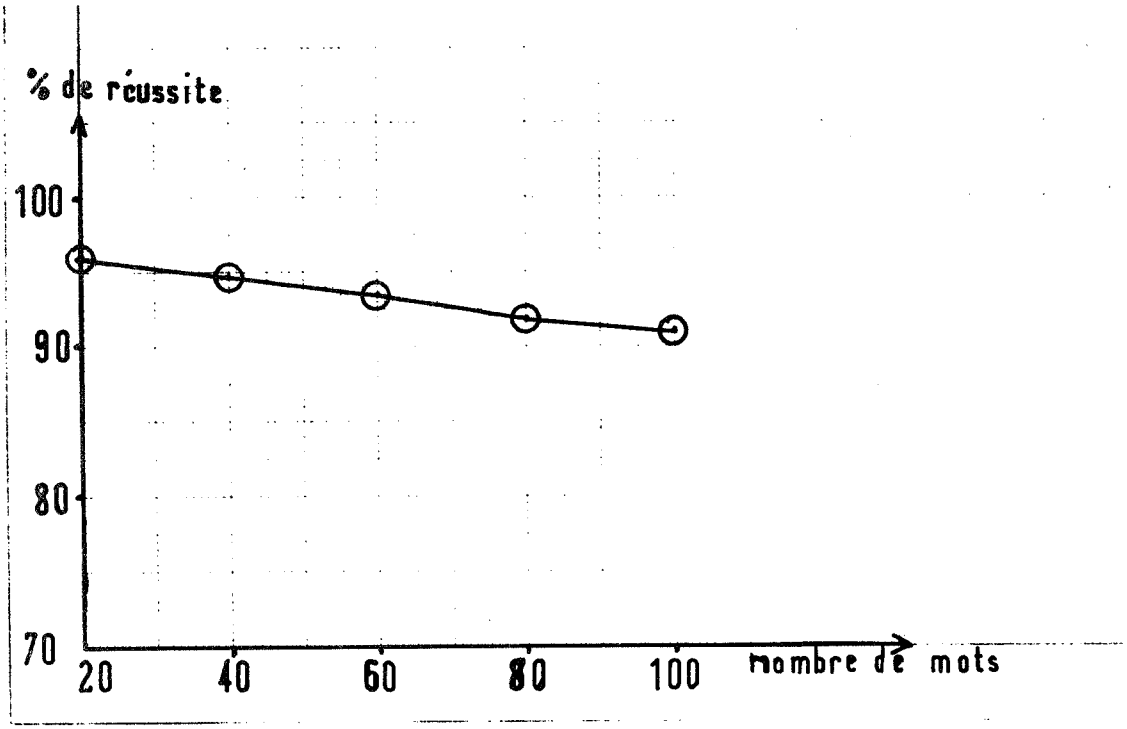
pas suivant. La note finale de ressemblance sera la moyenne des notes de ressemblance entre deux vecteurs "delta".



$$\text{note finale} = \frac{N1 + N2 + N3 + N4}{4}$$

Résultats :

Un locuteur a prononcé les 100 premiers mots de la liste CNET d'intelligibilité du téléphone. Ces 100 mots ont été catalogués comme dictionnaire. Le même locuteur a de nouveau prononcé ces 100 mots dans des conditions analogues. Pour chacun d'eux on a fait exécuter le programme de reconnaissance. Les résultats ont été dépouillés sur divers sous-ensembles du dictionnaire, de façon à caractériser l'évolution du système en fonction de l'étendue du vocabulaire. Le taux de reconnaissance varie de 96% pour 20 mots à 91% pour 100 mots



Un résultat comparable a été enregistré (94%) sur un autre vocabulaire (les 26 prénoms du code d'épellation téléphonique) lu une fois en apprentissage et deux fois en reconnaissance par le même locuteur.

Avec le même dictionnaire le système a été testé par d'autres locuteurs, conduisant à des résultats divers, compris entre 55 et 85% selon le locuteur. Mais ces chiffres n'ont qu'une faible signification, dans la mesure où aucun processus d'adaptation au locuteur n'a été mis en oeuvre. Nos recherches sur cet aspect du problème, que nous considérons comme très important, sont menées en parallèle (3).

Conclusion

Le système, tel qu'il existe actuellement, n'est qu'une première ébauche plus dirigée vers la recherche fondamentale que vers l'application. La majorité des erreurs observées relèvent plus de la technique (exploitation de la fonction de stabilité, processus de comparaison dynamique) que des paramètres utilisés. Nous poursuivons donc dans cette direction, en cherchant à rendre le processus insensible aux variations de la voix, pour un locuteur et pour plusieurs, et à réduire l'apprentissage au strict minimum.

Références

(1) E. LEIPP, J.S. LIENARD, M. CASTELLENGO, J. SAPALY, D. TEIL, A. CALINET, M. MLOUKA
Le colloque sur la parole - Bulletin n° 53 du Groupe d'Acoustique Musicale de l'Université Paris VI - Janvier 1971

(2) M. MLOUKA et J.S. LIENARD
Reconnaissance de mots caractérisés soit par leurs vecteurs de stabilité, soit par leurs vecteurs de transition - Eight International Congress on Acoustics - London July 1974

(3) J.S. LIENARD et M. MLOUKA
Normalisation fréquentielle de la parole - 4e Journées d'Etude sur la Parole (GALF) - Mai 1973

(4) J.S. LIENARD, M. CASTELLENGO, E. LEIPP, M. MLOUKA, G. RENARD, J. SAPALY, D. TEIL
Quelques idées directrices en reconnaissance automatique de la parole - Revue "Automatisme" n° 3 - Mars 1973

Table des Matières

du Volume I

---o---

Imprimé en France
Dépôt légal : 3° trimestre 1974

Service de Reprographie
C.N.R.S.
Gif