Bibliothèque
Institut de la
Communication Parlée

Groupement des Acousticiens de Langue Française

9èmes JOURNEES D'ETUDES SUR LA PAROLE

du Groupe de la

"COMMUNICATION PARLEE"

VOL. I



Journées d'Etudes des 31 mai au 2 juin 1978 au
Centre National d'Etudes des Télécommunications
à LANNION (22300)

Groupement des Acousticiens de Langue Française

ACTES
des

9èmes JOURNEES D'ETUDES SUR LA PAROLE
du
Groupe de la
"COMMUNICATION PARLEE"

LANNION

31 mai - 2 juin 1978

cnet
BP 40
22301 LANNION Cedex
France

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

Institut de Phonétique
Inventaire n° 2084
Cote n° A/JEP 9/1

Les 9èmes JOURNEES D'ETUDES SUR LA PAROLE qui se déroulent les 31 mai, 1er et 2 juin 1978 au Centre National d'Etudes des Télécommunications de LANNION sont organisées par les départements " Etudes et Techniques d'Acoustique " et " Services Spéciaux du Téléphone " du groupement " Distribution, Acoustique et Services Spéciaux ", sous le patronage du Groupement des Acousticiens de Langue Française.

Le travail est réparti au cours de trois demi-journées consacrées aux thèmes suivants :

1. LE TRAITEMENT PERCEPTIF DES INDICES ACOUSTIQUES ET DES TRAITS PHONETIQUES :

Président	:	Max	WAJSKOP
Rapporteurs	:	Christel	SORIN
		Willy	SERNICLAES

2. DESCRIPTION ARTICULATOIRE ET ACOUSTIQUE DES INDICES ET DES TRAITS

Président	:	René	CARRE
Rapporteurs	:	Mario	ROSSI
		Bernard	GUERIN

3. ANALYSE DES INDICES ET DETECTION AUTOMATIQUE DES TRAITS

Président	:	Guy	PERENNOU
Rapporteurs	:	Philippe	MARTIN
		Jean-Paul	HATON

Deux tables rondes complètent ces Journées d'Etudes ; elles sont consacrées aux sujets suivants :

Table ronde n° 1 : "Stimulus de Laboratoire et parole naturelle"

(Peut-on transposer à la parole naturelle les résultats, d'expériences réalisées à partir de stimuli artificiels ?)

Animateurs	:	Max	WAJSKOP
		Elisabeth	LHOTE

Table ronde n° 2 : "Evaluation de la qualité et de l'intelligibilité de la Parole synthétique"

Animateurs	:	Jacques	GENIN
		Joël	LEROUX
		Daniel	TEIL

.../...

Afin de donner à ces rencontres le caractère de journées de travail et de réflexion sur un thème bien défini, le Bureau du Groupe a demandé à des rapporteurs :

- de faire un exposé de synthèse sur la base des communications proposées ;

- de susciter des discussions en demandant aux auteurs des éclaircissements ou des compléments d'information et en critiquant les thèses ou les méthodes développées dans les Communications.

Les auteurs sont conviés, sans avoir à répéter le texte publié dans les Actes, à compléter, illustrer ou justifier leur méthode ou leurs thèses.

Ce volume contient le texte complet des Communications présentées.

Les textes des rapporteurs seront publiés ultérieurement dans un prochain numéro de la "Revue d'Acoustique" éditée par le G.A.L.F.

Comité organisateur :

Raymond	DESCOUT
Claire	ISSLER
Louis	LE MARTRET
Marie-Claude	MAISONNEUVE
Guy	MERCIER
Pierre	STEPHAN
Roland	VIVES

The 9èmes JOURNEES D'ETUDES SUR LA PAROLE sponsored by the Groupement des Acousticiens de Langue Française is held at the Centre National d'Etudes des Télécommunications, LANNION, on May 31, June 1 and 2 1978. This conference is organized by the members of the 2 departments : "Etudes et Techniques d'Acoustique", and "Services Spéciaux du Téléphone" both from the "Distribution, Acoustique et Services Spéciaux", division of the C.N.E.T.

The 9èmes JOURNEES D'ETUDES SUR LA PAROLE consists of two types of activities : three consecutive half-day long sessions on three different topics :

1. PERCEPTUAL TREATMENT OF ACOUSTIC CUES AND PHONETIC FEATURES:

Chairman	:	Max	WAJSKOP
Reporters	:	Christel	SORIN
		Willy	SERNICLAES

2. ARTICULATORY AND ACOUSTICAL REPRESENTATION OF ACOUSTIC CUES AND PHONETIC FEATURES:

Chairman	:	René	CARRE
Reporters	:	Mario	ROSSI
		Bernard	GUERIN

3. AUTOMATIC DETECTION AND ANALYSIS OF ACOUSTIC CUES AND PHONETIC FEATURES:

Chairman	:	Guy	PERENNOU
Reporters	:	Philippe	MARTIN
		Jean-Paul	HATON

Two workshops :

Workshop No 1 : "Laboratory stimulus and natural speech"

(What can we infere on natural speech from results obtained on laboratory stimuli ?)

Chairmen	:	Max	WAJSKOP
		Elisabeth	LHOTE

Workshop No 2 : "Evaluation of synthetic speech : Intelligibility and naturalness"

Chairmen	:	Jacques	GENIN
		Joël	LEROUX
		Daniel	TEIL

.../...

In order to insure deeper involvment of all attendants to sessions, the conference record is preprinted and circulated to each participant before the conference. During the sessions, there will be no formal presentation by authors. However, the chairmen are kindly requested :

- to draw an overview of all materials submitted ;

- to initiate discussion with authors for further information, critical comparison between results, hypothesis and theories.

A short period of time is allowed to each author so he can complete or clarify what he put forward in this paper by using, for example, audiovisual support.

Chairmen's reports will be published in the "Revue d'Acoustique", a periodic publication of the Groupement des Acousticiens de Langue Française (G.A.L.F.) in coming month issue.

Committee of organization :

Raymond	DESCOUT
Claire	ISSLER
Louis	LE MARTRET
Marie-Claude	MAISONNEUVE
Guy	MERCIER
Pierre	STEPHAN
Roland	VIVES

TABLE DES MATIERES

Séance consacrée au thème 1 :

"LE TRAITEMENT PERCEPTIF DES INDICES ACOUSTIQUES ET DES TRAITS PHONETIQUES".

p.

B. DELGUTTE (M.I.T - Cambridge) 11

Codage des changements rapides d'intensité dans le nerf auditif : expériences avec des sons purs.

P. MARTIN (University of Toronto) 21

Perception des séquences de contours prosodiques de phrases synthétisées.

E. VIVALDA, P.M. BERTINETTO (C.S.E.L.T. - Turin) 31

Perception de la quantité vocalique et consonantique en italien, par analyse-synthèse de la voix.

P. BOVET, M. ROSSI (Inst. de Phonétique - Aix en Provence) 43

Etude comparée de la sensibilité différentielle à la durée avec un son pur et avec une voyelle.

C. SORIN (CNET - Lannion) 57

Seuils différentiels et perception de modifications d'intensité dans la parole continue.

M. CHAFCOULOFF, A. DI CRISTO (Inst. de phonétique - Aix en Provence) 69

Les indices acoustiques et perceptuels des consonnes constrictives du français ; application à la synthèse.

R. BECKMANS (Inst. de Phonétique - Bruxelles) 83

Structuration perceptive de deux indices acoustiques et perception catégorielle.

Séance consacrée au thème 2 :

"DESCRIPTION ARTICULATOIRE ET ACOUSTIQUE DES INDICES ET DES TRAITS".

p.

J. CAELEN, G. PERENNOU (CERFIA - Toulouse) 95

Indices et traits acoustiques dans un système de reconnaissance de la parole continue : quelques résultats.

M. ROSSI, C. LE CORRE, G. MERCIER (Inst. de Phonétique - Aix en Provence CNET - Lannion) 105

Indices de détection de formants sur analyse spectrale par canaux.

M.T. ROTH (CNET - Lannion) 115

Etude des variations acoustiques de la voyelle dans les monosyllabes en français.

G. KONOPCZYNSKI, S. PRUDHAM, R. VARGAS, C. VITERI (Lab. de Phonétique - Besançon) 141

Les indices de l'accent tonique et leur hiérarchie : application à l'espagnol.

S. BARTH, S. CHULLIAT (Inst. de jeunes sourds - Cognin) 149

Modifications inter-locuteurs de l'échelle formantique

C. ABRY, L.J. BOË (Inst. de phonétique - Grenoble) 157

Essai d'analyse phonologique des indices du voisement.

B. GUERIN, L.J. BOË (ENSERG - Grenoble, Inst. de Phonétique - Grenoble) 167

Etude d'un indice acoustique des voyelles : la puissance intrinsèque.

R. DESCOUT, L.J. BOË, C. ABRY (Inst. de Phonétique - Grenoble, CNET - Lannion) 177

Labialité vocalique et labialité consonnantique en français ; premiers résultats.

S. MAEDA (CNET - Lannion) 191

Une analyse statistique sur les positions de la langue : étude préliminaire sur les voyelles françaises.

N. THORSEN (Inst. de Phonétique - Copenhague) 201

Une explication simplifiée, en termes physique, des conséquences acoustiques des mouvements de la langue et des lèvres dans la production des voyelles.

Séance consacrée au thème 3 :

"ANALYSE DES INDICES ET DETECTION AUTOMATIQUE DES TRAITS". p.

J.A DREYFUS-GRAF (Genève) 213

Caractéristiques comparées de systèmes reconnaissant la parole.

G. PERENNOU, J. CAELEN (CERFIA - Toulouse) 223

Localisation des voyelles dans le plan (F1, F2) application à la reconnaissance de la parole.

T. DE GRAAF (Inst. des Sc. Phonétiques - Groningue) 233

Analyse de voyelles avec des méthodes digitales.

N. TIRANDAZ, C. BERGER - VACHON (Lab. de Physique Electronique - Lyon) 241

Analyse d'une méthode probabiliste appliquée à la reconnaissance des voyelles parlées.

G. CAELEN, G. MAURAND (UER Linguistique - Toulouse) 251

La durée des phonèmes et la reconnaissance des débuts et fins de mots : étude acoustique des structures homophoniques chez deux locuteurs.

A.C.M. RIETVELD, L. BOVES (Inst. Voor. Fonetiek-Nijmegen) 261

Détection automatique de syllabes accentuées en Néerlandais.

C. SANCHEZ, G. MESSENET, J.P. HATON (CRIN - Nancy) 271

Etude et utilisation des indices acoustiques et des traits pour la segmentation et la reconnaissance phonémique de la parole.

Y. RENE DE COTRET (Lab. de Physique - Liège) 277

Le modèle linéaire de production de la parole utilisé pour la reconnaissance automatique de voyelles non nasalisées de la langue française.

R. GUBRYNOWICZ, R. DE MORI, P. LAFACE (Ac. Polon. des Sciences - Varsovie, Univers. de Turin - Turin). 287

La description au niveau acoustique des consonnes nasales prononcées dans un discours continu.

P. ALINAT (Thomson - C.S.F - Cagnes sur mer) 297

Etude du trait permettant de distinguer entre les 3 classes de consonnes explosives PB, TD, KG.

M. LAMOTTE, H. LEM, C. VIGNERON, M.J. VIGNERON 305

(Lab. d'automatique - Nancy)
Détection automatique de traits acoustiques, en temps réel, dans la parole continue.

M. BAUDRY, B. DUPEYRAT (CEN - Saclay) 313

Utilisation de méthodes syntaxiques pour la détection automatique des traits phonétiques en reconnaissance de la parole.

G. MERCIER (CNET - Lannion) 321

Evaluation des indices acoustiques utilisés dans l'analyseur phonétique du système KEAL.

THEME 1

LE TRAITEMENT PERCEPTIF DES INDICES ACOUSTIQUES ET DES TRAITS PHONETIQUES

9èmes JOURNEES D'ETUDE SUR LA PAROLE

LANNION 31 mai - 2 juin 1978

CODAGE DES CHANGEMENTS RAPIDES D'INTENSITE DANS LE NERF AUDITIF: EXPERIENCES AVEC DES SONS PURS

Bertrand DELGUTTE

Eaton-Peabody Laboratory of Auditory Physiology, Massachusetts Eye and Ear Infirmary, Boston, Massachusetts, and Research Laboratory of Electronics, MIT, Cambridge, Massachusetts, U.S.A.

.**RESUME** Ces expériences font partie d'une étude électrophysiologique du codage des sons de la parole dans le nerf auditif. Le principe de cette étude consiste à utiliser des stimuli qui ont certaines propriétés de la parole, mais qui demeurent assez simples pour permettre l'interprétation des résultats en fonction de ce que l'on connait déjà sur le codage des sons simples. On introduira progressivement dans les expériences des stimuli plus proches de la parole. En raison de l'importance des changements rapides d'intensité dans la parole, nous avons choisi comme stimuli des sons purs (analogues à un formant) ayant une attaque et une chute abruptes. On utilise des chats pour ces expériences afin de pouvoir mesurer la réponse de fibres individuelles du nerf auditif.

A l'attaque du son pur, le taux de déclenchement des potentiels d'action pour les fibres stimulées atteint rapidement un maximum, puis décroît graduellement et approche enfin une valeur stationnaire. Cette décroissance graduelle du taux de déclenchement a deux constantes de temps: l'une, rapide, de l'ordre d'une période de la voix; l'autre, lente, de l'ordre de la durée d'un élément phonétique. La dynamique de certaines fibres est plus étendue pour une attaque abrupte que pour un son stationnaire.

Dans certaines expériences, on a présenté un second son pur quelques msec après la chute abrupte du premier. La réponse des fibres stimulées au second son est plus faible en présence du premier son qu'en son absence: il y a masquage a posteriori. Dans la parole, le masquage pourrait modifier la représentation de l'enveloppe spectrale dans le nerf auditif pour les sons qui suivent un changement abrupt de spectre.

RESPONSES OF SINGLE AUDITORY-NERVE FIBERS TO TONE BURSTS*

Bertrand DELGUTTE, Eaton-Peabody Laboratory of Auditory Physiology, Massachusetts Eye and Ear Infirmary, Boston, Massachusetts, and Research Laboratory of Electronics, MIT, Cambridge, Massachusetts, U.S.A.

SUMMARY The present experiments are part of an electrophysiological study on the coding of speech-like sounds in the auditory nerve. Tone bursts with abrupt onsets and offsets were chosen as stimuli for the initial experiments. These stimuli have properties which represent a first approximation to the abrupt changes in amplitude that occur in speech.

The electrical activity of single auditory-nerve fibers in response to 200 msec tone bursts with a 2.5 msec rise-fall time was recorded with microelectrodes in anesthetized cats. The level and frequency of the tone burst were varied within limits typical for speech. Poststimulus time histograms of the response to 10 levels of tone bursts are shown in Fig. 3. After the stimulus onset, the firing rate quickly reaches a maximum, then decreases gradually, approaching a steady-state value after 100-150 msec. This gradual decrease in rate, or "adaptation," can be characterized by the superposition of a fast exponential decay having a time constant of 3-6 msec and a slower exponential decay with a 30-45 msec time-constant. For some units (Fig. 4A), the ratio of the maximum firing rate at onset to the steady-state rate is virtually independent of stimulus level. For other units (Fig. 4B), the ratio increases with level, and the dynamic range can be wider for abrupt onsets than for steady tones.

The adaptation time constants are independent of stimulus frequency and fiber characteristic frequency (CF). Thus, the values of the time constants might be applicable to sounds with complex spectra such as speech, without regard to the actual frequencies involved. The fast adaptation time constant is on the order of a voice period, and one would expect adaptation to interact with the synchronization of spikes to the fundamental frequency of voice. The slow adaptation time constant is on the order of the duration of a phonetic segment, and syllables which have a stop consonant in initial position should produce a prominent adaptation in single auditory-nerve fiber responses.

In other experiments, the 200 msec tone burst was followed after a variable time delay by a second "probe" tone burst, 20 msec in duration, 20 dB above threshold at CF (Fig. 5). The decrease in response to the probe due to the presence of the first "masker" tone burst was measured for various time delays and masker levels (Fig. 6). For a zero time delay, a masker 10-20 dB above threshold reduced the probe response by 50%. At high masker levels, the probe response could be completely suppressed. When the time delay was varied, the time constant for masking was found to be larger than the adaptation time constant, and to increase with masker level. In speech, forward masking by a prolonged sound (e.g. vowel) followed by an abrupt spectral change could modify the representation of the spectral envelope for the masked sounds in the auditory nerve.

*This work was supported by U.S. Public Health Service grants 5 P01 NS13126 and NS 04332 .

INTRODUCTION

Jusqu' à présent, les études sur la perception de la parole ont été fondées sur un modèle simplifié du système auditif périphérique selon lequel l'oreille opère comme un banc de filtres sur le stimulus; le signal d'entrée au système nerveux est une version codée des variations de l'énergie à la sortie de chaque filtre en fonction du temps. D'après ce modèle, le codage de la parole dans le nerf auditif ressemble à un spectrogramme. Durant les quinze dernières années, les études électrophysiologiques ont montré que, si ce modèle demeure une approximation raisonnable, il est nettement insuffisant pour décrire le codage de sons complexes tels que ceux de la parole. Pour produire un modèle plus réaliste, il est nécessaire de réunir des données systématiques sur le traitement des sons de la parole par le système auditif périphérique.

Traditionnellement, on a utilisé des méthodes indirectes pour obtenir des informations sur le traitement de la parole par l'oreille, par exemple en généralisant à la parole des résultats psychoacoustiques sur des sons simples, ou bien en faisant des expériences perceptives avec de la parole naturelle ou synthétique. Une autre méthode - l'enregistrement électrophysiologique de la réponse du nerf auditif à des sons analogues à ceux de la parole - n'a pas été utilisée, bien qu'elle ait le potentiel de donner les résultats les plus directs et les plus précis sur le codage des indices acoustiques. Il est maintenant possible d'enregistrer à l'aide de microélectrodes l'activité électrique de fibres individuelles du nerf auditif chez le chat ou d'autre mammifères. Cette approche se justifie par les raisons suivantes: (1) il y a une grande similarité entre l'organisation de la cochlée du chat et celle de l'homme; (2) toute l'information auditive parvenant au cerveau doit être codée dans le nerf auditif; donc, si un indice acoustique n'est pas présent au niveau du nerf auditif, le système nerveux central ne peut pas l'utiliser pour discriminer les sons de la parole; (3) on a déjà étudié systématiquement la réponse du nerf à des sons simples (clics, sons purs continus, bruit blanc); ces résultats donnent une première idée de la façon dont les sons complexes sont codés.

Le succès de ces expériences électrophysiologiques sur le codage de la parole dépend étroitement du choix adéquat des stimuli. Techniquement, il serait possible de mesurer la réponse du nerf à la parole naturelle. Cependant, l'interprétation de tels résultats serait obscurcie par des intéractions complexes des differents indices acoustiques présents dans le stimulus, et il serait difficile de comprendre comment chacun des indices est codé. Il est préférable d'utiliser des stimuli qui possèdent une propriété élémentaire de certains sons de la parole (par exemple une attaque abrupte), mais dont on puisse interpréter le codage par comparaison avec des résultats pour des stimuli plus simples. Par exemple, on peut comparer la réponse du nerf à des sons purs continus (JOHNSON, 1974) avec la réponse à des sons purs ayant une attaque abrupte. Certains chercheurs ont fait quelques pas dans cette direction (SMITH et al., 1975; SMITH, 1977; HARRIS, 1977).

Etant donné l'importance des changements abrupts d'intensité dans la parole nous avons entrepris une série d'expériences sur des sons purs ayant une attaque et une chute abruptes. Ces expériences sont décrites dans la section II. Dans la section III, nous reportons une expérience sur l'influence d'un son pur ayant une chute abrupte sur la réponse du nerf à un autre son présenté quelques millisecondes après la chute du premier ("masquage a posteriori"). La section I est un résumé des techniques d'analyse des réponses électrophysiologiques, et des résultats élémentaires sur le codage des sons simples. Pour un lecteur qui ne connait pas bien les enregistrements électrophysiologiques, cette section est essentielle a la compréhension des expériences reportées dans les sections II et III.

I TECHNIQUES ET RESULTATS DES ENREGISTREMENTS ELECTROPHYSIOLOGIQUES

A. Codage des sons simples (Pour une introduction, voir KIANG, 1975; KIANG et al., 1972; pour un traitement plus complet, voir KIANG et al., 1965; KIANG et al., 1974; JOHNSON, 1974). Une microélectrode contactant une fibre du nerf auditif enregistre une série d'impulsions, ou potentiels d'action (PA), qui durent moins de 1 msec, et qui ont toutes à peu près la même amplitude et la même forme d'onde. Seule la suite des instants où ces impulsions se produisent transmet de l'information sur le stimulus. En l'absence de son, on observe l'activité spontanée de la fibre, qui est constituée de PA se produisant à des instants aléatoires avec une fréquence moyenne (ou taux d'activité spontanée) de 0 à 150 impulsions/ sec selon la fibre.

Si on présente à l'animal un son pur de fréquence et d'intensité appropriées, le taux de déclenchement moyen des PA s'élève au-dessus du taux d'activité spontanée. L'intensité à laquelle un son pur de fréquence donnée provoque une augmentation tout juste discernable du taux moyen de déclenchement des PA est le seuil de la fibre à cette fréquence. La courbe du seuil en fonction de la fréquence du stimulus (courbe seuil-fréquence) divise le plan fréquences/intensités en deux régions: pour les sons purs situés en dessous de cette courbe, la fibre ne répond pas (le taux de déclenchement des PA est égal au taux spontané); pour les sons purs situés au-dessus de la courbe, la fibre répond. La fréquence à laquelle la fibre a le seuil le plus bas (c'est-à-dire la fréquence où la fibre est la plus sensible) est la fréquence caractéristique (FC) de la fibre. La Fig. 1 montre une série de courbes seuil-fréquence pour des fibres de différentes FC. Toutes les courbes ont une "pointe" étroite centrée à la FC. Ainsi, aux faibles intensités, toute fibre se comporte comme un filtre passe-bande étroit. Toutefois, les courbes des fibres ayant une FC supérieure à 3 kHz ont aussi une large "queue" qui s'étend loin de la FC dans les basses fréquences. Ainsi, pour les intensités élevées, les fibres ayant une FC au-dessus de 3 kHz répondent aussi aux sons de basse fréquence. Les sons de la parole qui n'ont que des composantes de basse fréquence stimulent les fibres ayant une FC élevée aux intensités supérieures à 60 dB.

Des fibres ayant la même FC ont des courbes seuil-fréquence qui se ressemblent fort, et on peut prévoir dans une large mesure la réponse à un son pur arbitraire à partir de la seule connaissance de la FC. (Cependant, le seuil à la FC peut varier sur un intervalle de 40 dB d'une fibre à l'autre.) Si on augmente l'intensité d'un son pur entre 0 et environ 30 dB au-dessus du seuil, le taux moyen de déclenchement des PA augmente aussi; au delà, le taux se stabilise: la dynamique de la plupart des fibres n'est que de l'ordre de 30 dB. Un autre aspect de la réponse à des sons purs est que, pour des fréquences inférieures à 4-5 kHz les PA tendent à se produire à un instant bien précis de la période du son pur: la réponse de la fibre est synchronisée avec la fréquence du stimulus.

En conclusion, l'information relative au stimulus est codée de deux facons dans le nerf auditif. Premièrement, il y a un codage fibre à fibre: étant donné que les fibres se comportent comme des filtres, chacune transmet une partie de l'information sur le spectre du stimulus. Pour la parole, l'information relative à l'envelope spectrale est vraisemblablement codée de cette manière dans une large mesure. Deuxièmement, il y a un codage temporel qui se manifeste de deux facons: (1) les variations d'intensité se traduisent par des variations du taux de déclenchement des PA pour certaines fibres; (2) la réponse des fibres est synchronisée avec les fréquences présentes dans le stimulus. Pour la parole, il est probable que l'information relative à l'intensité, au voisement, et à la fréquence fondamentale est codée temporellement.

B. Histogramme relatif au temps du stimulus (GERSTEIN et al., 1960). La Fig. 2A montre la réponse d'une fibre à 6 présentations d'un même son pur de 200 msec suivi par un

silence de 200 msec. Chaque PA est représenté par une barre verticale. Les PA ne se produisent pas aux mêmes instants pour les différentes présentations du stimulus, mais certaines régularités sont visibles: à l'attaque du stimulus, le taux de déclenchement des PA est maximal, puis il diminue graduellement pendant les 100 premières msec du stimulus; après la cessation du stimulus, le taux de déclenchement des PA est pratiquement nul, puis il remonte progressivement vers le taux d'activité spontanée. On peut faire apparaître ces régularités d'une facon plus précise en calculant un histogramme des PA relatif au temps qui suit l'attaque du stimulus (histogramme RTS). Pour cela (1) on présente le même stimulus de nombreuses fois, (2) on divise le temps qui suit l'attaque du stimulus en un grand nombre de petits intervalles, et (3) on fait la somme du nombre des PA qui se sont produits dans chaque intervalle de temps pour toutes les présentations du stimulus. La Fig. 2B montre le résultat de ce calcul pour 300 présentations du stimulus, et pour 200 intervalles de 2 msec (la ligne pointillée représente le taux d'activité spontanée). Cette technique permet d'estimer avec la résolution temporelle choisie les variations du taux instantané de déclenchement des PA en fonction du temps pour un stimulus arbitraire.

II REPONSE DU NERF AUDITIF A DES SONS PURS AYANT UNE ATTAQUE ABRUPTE

A. Le problème. La Fig. 2B montre un histogramme RTS en réponse à un son pur de 200 msec suivi par un silence de 200 msec. La décroissance graduelle du taux de déclenchement des PA après l'attaque du stimulus est appelée adaptation. La remontée du taux de déclenchement après la cessation du stimulus est appelée rétablissement de l'activité spontanée. On observe adaptation et rétablissement non seulement pour les sons purs, mais aussi pour les sons complexes comme la parole. En fait, chaque fois qu'un changement abrupt d'intensité se produit dans la parole, on peut voir l'effet de l'adaptation et du rétablissement dans la réponse de certaines fibres. De plus, étant donné qu'une fibre se comporte comme un filtre passe-bande, un changement abrupt dans le spectre du stimulus peut se traduire pour la fibre comme un changement d'intensité. Finalement, pour les sons voisés, on peut considérer chaque période du larynx comme un interrupteur qui module l'énergie des bandes formantiques. Si l'adaptation avait une constante de temps comparable à la période de la voix, on s'attendrait à ce qu'elle interagisse avec la modulation du signal de parole. Ainsi, l'adaptation a peut-être un effet sur le codage des voyelles, et sur celui des consonnes (qui se caractérisent souvent par un changement abrupt de spectre et/ou d'intensité).

Pour étudier l'adaptation, nous avons calculé des histogrammes RTS en réponse à des sons purs de 200 msec ayant une attaque linéaire de 2,5 msec. On a calculé ces histogrammes avec une résolution de 1 msec, puis on les a lissés par convolution avec une fenêtre de Hamming de 6 msec. Pour chaque fibre, on a fait varier l'intensité et la fréquence du son pur dans un domaine caractéristique à la parole. On a étudié ainsi plus de 200 fibres de FC variant entre 200 Hz et 10 kHz.

B. Résultats. La Fig. 3 montre une série d'histogrammes RTS pour un son pur à la FC de la fibre. L'intensité du stimulus varie de 6 dB à 60 dB au-dessus du seuil. A l'exception des plus basses intensités, on peut décrire chacun de ces histogrammes par une superposition de trois composantes: (1) une composante stationnaire qui dure le temps du stimulus, (2) une composante d'adaptation rapide qui dure environ 15-20 msec après le pic de l'histogramme, (3) une composante d'adaptation lente qui dure 100-150 msec. On peut caractériser la composante rapide et la composante lente par des exponentielles ayant respectivement une constante de temps de 3-6 msec et 30-45 msec. Toutefois, pour les intensités moins de 10 dB au-dessus du seuil (Fig. 3), une seule composante d'adaptation ayant une constante de temps de 10-15 msec est suffisante. Il semble que, sauf aux faibles intensités les valeurs des constantes de temps sont à peu près indépendantes de l'intensité et de la fréquence du stimulus, et de la FC de la fibre.

La Fig. 4A montre la variation du taux de déclenchement des PA (1) à l'attaque du son et (2) en régime stationnaire, en fonction de l'intensité du son. Le stimulus est à la FC de la fibre. Le seuil pour la composante stationnaire est égal au seuil à l'attaque. Quand on augmente l'intensité, les deux courbes se séparent, mais croissent et saturent d'une facon similaire: le rapport du taux à l'attaque au taux stationnaire demeure constant et égal à 4. Les deux courbes saturent à la même intensité. La Fig. 4B montre les mêmes courbes pour une autre fibre. Le taux de régime stationnaire sature à 20 dB au-dessus du seuil, mais le taux de déclenchement à l'attaque du son continue d'augmenter sur 20 dB. Le rapport du taux d'attaque au taux stationnaire augmente avec l'intensité, et atteint la valeur 8. Les fibres que nous avons étudiées se distribuent entre les deux extrêmes représentés par les Fig. 4A et 4B. Le rapport du taux d'attaque au taux stationnaire varie de fibre à fibre entre 2,5 et 10. Il peut augmenter ou rester constant avec l'intensité, mais ne diminue jamais.

Quand on présente un son pur dont la fréquence est située dans une région où la courbe seuil-fréquence a un fort gradient, on fait les observations suivantes: (1) le seuil à l'attaque du stimulus est plus bas que le seuil en régime stationnaire; (2) aux basses intensités, l'histogramme RTS se limite à un pic bref à l'attaque du son; (3) quand on élève l'intensité, la composante stationnaire et les composantes d'adaptation apparaissent, et l'histogramme RTS retrouve la forme normale. Ces observations s'expliquent sans doute par le fait que, à cause de la forme de la courbe seuil-fréquence, le spectre étalé à l'attaque du son pur excite la fibre plus efficacement que la partie stationnaire du son pur. Bien que peu de fibres soient affectées par ce phénomène, pour chaque présentation d'un son à bande étroite il y a un petit nombre de fibres qui ne répondent apparemment qu'à l'attaque du son.

C. Implications pour la parole. Trois aspects de ces résultats sont importants pour la parole. Premièrement, le fait que les paramètres de l'adaptation (constantes de temps, rapport du taux d'attaque au taux stationnaire) ne dépendent pas de la FC de la fibre, ni de la fréquence du son, suggère que nos résultats sur les sons purs sont applicables à des sons ayant un spectre complexe, quelles que soient les fréquences présentes dans le stimulus. Deuxièmement, l'adaptation comprend une composante rapide et une composante lente dont les effets sur le codage de la parole se manifesteront à des échelles de temps différentes. La composante lente a une constante de temps de l'ordre de la durée d'un élément phonétique. Pour les syllabes ayant une attaque abrupte (par exemple celles qui ont une occlusive ou une affriquée en position initiale), on peut s'attendre à ce que l'adaptation apparaisse nettement dans la réponse de certaines fibres. Pour les syllabes ayant une attaque graduelle (par exemple celles qui commencent par une fricative, une semi-voyelle ou une voyelle) l'adaptation devrait être moins proéminente dans la réponse des fibres. La composante rapide de l'adaptation a une constante de temps de l'ordre de la période d'un son voisé. Ainsi, on peut prédire que l'adaptation influe sur le codage de chaque période de la parole. Cependant, il n'est pas possible de décrire précisément cette influence, car un son pur est une mauvaise approximation d'un son voisé du point de vue de la forme d'onde, et il serait utile de faire des expériences avec des sons plus réalistes tels que voyelles à un formant. Troisièmement, certaines fibres ont une dynamique plus étendue pour les sons ayant une attaque abrupte que pour les sons stationnaires. Pour ces fibres, le taux de déclenchement des PA peut être dix fois plus élevé à l'attaque du stimulus qu'en régime stationnaire. Ce résultat suggère que le système auditif périphérique traite les attaques abruptes d'une facon privilégiée. On peut étendre cette conclusion à tout changement abrupt d'intensité, puisque SMITH et al. (1975) ont montré que le changement instantané du taux de déclenchement des PA en réponse à un petit accroissement ou décrément d'intensité demeure le même quand la fibre a subi une adaptation et quand elle est rétablie.

III MASQUAGE A POSTERIORI

A. Le problème. Après la chute abrupte d'un son pur, on observe pour les fibres stimulées une dépression du taux de déclenchement des PA en dessous du taux spontané, puis un rétablissement graduel (Fig. 2B). La dépression n'affecte pas seulement l'activité spontanée, mais aussi la réponse aux sons qui suivent: si on présente un stimulus-sonde quelques msec après la chute du son pur (Fig. 5), la réponse à cette sonde est réduite par rapport à la valeur qu'elle aurait en l'absence du son pur; cet effet est analogue au masquage a posteriori des expériences de psychoacoustique, mais il y a d'importantes différences. En psychoacoustique, on dit qu'il y a masquage quand la sonde ne peut être détectée. Ici, on peut simplement dire que la réponse d'une fibre à la sonde est modifiée par la présence du "masqueur." Au bout d'un certain temps après la chute du masqueur, la fibre est rétablie: le masqueur n'affecte plus la réponse de la fibre à la sonde. Le masquage est beaucoup moins fort pour un masqueur de courte durée (moins de 20 msec) que pour un masqueur de plus de 100-150 msec: il faut que la fibre ait le temps de subir une adaptation pendant la durée du masqueur pour qu'il y ait masquage (SMITH, 1977).

Les sons de la parole ne sont pas prononcés isolément, et on peut s'attendre à ce qu'un son prolongé influe sur le codage des sons qui le suivent. Par exemple, pour une syllabe voyelle-plosive, le 2ème formant de la voyelle est analogue à un son pur masqueur qui pourrait affecter la réponse du nerf au bruit de frication de la plosive après les quelques msec du silence d'occlusion. Dans le cas particulier où l'intervalle de silence entre le masqueur et la sonde est nul, notre stimulus à deux sons purs devient analogue à un changement abrupt d'intensité et/ou de spectre.

Dans ces expériences, la sonde est toujours un son pur de 20 msec à la FC de la fibre, 20 dB au-dessus du seuil (Fig. 5). Le masqueur est le même son pur que celui de la section II. On fait varier l'intensité et la fréquence du masqueur, ainsi que la durée de l'intervalle de silence. Pour chaque valeur de ces paramètres, on présente le stimulus 80 fois, et on mesure le nombre total de PA qui se sont produits pendant la durée de la sonde. On appellera ce nombre "réponse à la sonde."

B. Résultats. La Fig. 6 montre la réponse d'une fibre à la sonde (en pourcentage de la réponse en l'absence du masqueur) en fonction de l'intensité du masqueur pour 4 valeurs de l'intervalle de silence (de 0 à 120 msec). Pour un masqueur en dessous du seuil, la réponse à la sonde reste égale à la réponse en l'absence du masqueur (100%). Si on augmente l'intensité du masqueur au-dessus du seuil, la réponse à la sonde décroît d'abord rapidement sur environ 20 dB, puis elle tend à se stabiliser aux intensités élevées. En général, pour un silence de 0 msec, la réponse à la sonde est réduite de moitié quand le masqueur est à 20 dB au-dessus du seuil; pour un masqueur à 40 dB au-dessus du seuil, elle n'est plus que 0-25%. Ainsi, le masquage peut supprimer complètement la réponse à la sonde.

Pour mieux voir comment le masquage dépend de l'intervalle de silence, nous avons reporté sur la Fig. 7 les points de la Fig. 6 en fonction de la durée du silence, pour 6 valeurs de l'intensité du masqueur (de 12 dB à 42 dB au-dessus du seuil). L'ordonnée est la diminution (en %) de la réponse à la sonde due au masqueur, c'est a dire 100% moins l'ordonnée de la Fig. 6. Chacune des 6 dourbes est une approximation exponentielle à la decroissance du masquage avec le temps pour une intensité du masqueur. Ces approximations ne sont pas très satisfaisantes, mais elles donnent une idée de la "constante de temps" du masquage (voir la légende de la Fig. 7). On voit que la constante de temps augmente avec l'intensité du masqueur. Pour un masqueur à 20 dB au-dessus du seuil, elle est en général de l'ordre de 50-120 msec; à 40 dB, elle atteint 100-200 msec. Ainsi, plus une fibre a reçu une stimulation intense, plus elle met longtemps à se rétablir. Notons aussi

que les constantes de temps du masquage sont beaucoup plus grandes que les constantes de temps pour l'adaptation. Les constantes de temps du masquage, ainsi que la sensibilité d'une fibre au masquage semblent être indépendantes de la FC de la fibre.

Jusqu'ici, nous n'avons considéré que la réponse totale à la sonde. En calculant un histogramme RTS, on peut aussi mesurer le taux instantané de déclenchement des PA en réponse à la sonde. En l'absence du masqueur, le taux est maximum à l'attaque de la sonde, puis décroît graduellement. Quand il y a masquage, le pic du taux de déclenchement à l'attaque de la sonde est moins proéminent, et pour un masqueur d'intensité élevée, il peut être supprimé. Ainsi, le masquage ne se traduit pas seulement par une diminution de la réponse à la sonde, mais aussi par une réduction apparente des effets de l'adaptation pour les sons masqués.

C. Implications pour la parole. Un son prolongé de la parole (par exemple une voyelle) qui a beaucoup d'énergie dans la région de son 2ème formant peut adapter les fibres ayant une FC autour du 2ème formant, ainsi que celles ayant une FC supérieure à 3 kHz (puisque la "queue" de leur courbe seuil-fréquence contient la région du 2ème formant). La réponse de ces fibres aux sons qui suivent subit donc un masquage qui peut durer pendant plusieurs éléments phonétiques, mais qui est surtout important pendant les 50 premières msec. Le masquage pourrait modifier le codage de l'enveloppe spectrale des sons masqués puisqu'il affecte certaines bandes de FC plus fortement que d'autres. Quand il y a un changement abrupt de spectre, ceci tendrait à accentuer le contraste entre le spectre du masqueur et le spectre des sons masqués dans la région des basses fréquences. Le masquage pourrait aussi affecter le codage des changements abrupts d'intensité pour les sons masqués, puisqu'il réduit parfois les effets apparents de l'adaptation.

CONCLUSION

Ces expériences montrent qu'avec des stimuli relativement simples, tels que des sons purs ayant une attaque abrupte, on peut étudier des phénomènes physiologiques qui affectent le codage des sons de la parole. Nos résultats sur l'adaptation et le masquage a posteriori sont souvent compatibles avec les modèles de l'adaptation (SIEBERT et al., 1968; SCHROEDER et al., 1974; SMITH et al., 1975). Cependant, aucun de ces modèles dans sa formulation stricte n'est en accord avec l'ensemble de nos résultats, et certains des résultats semblent être en contradiction avec tous ces modèles. L'utilisation de stimuli simples n'est pas suffisante pour étudier le codage de la parole, puisque les modèles de la réponse du nerf à des sons simples ne permettent pas toujours de prédire la réponse à des sons plus complexes. Ces considérations justifient notre approche de l'étude du codage de la parole par utilisation de stimuli progressivement plus complexes. Les implications fonctionnelles de certains de nos résultats ne sont pas toujours évidentes, mais on peut espérer qu'elles deviendront plus claires lorsque se développera une vue plus globale du codage de la parole dans le système auditif.

REFERENCES

GERSTEIN, G.L., KIANG, N.Y.S., 1960, An approach to the quantitative analysis of electrophysiological data from neurons. Biophys. J. 1, pp. 15-28.

HARRIS, D.M., 1977, Forward masking and recovery from short-term adaptation in single auditory-nerve fibers. Ph.D. Thesis, Northwestern University, Evanston, Ill.

JOHNSON, D.H., 1974, The response of single auditory-nerve fibers in the cat to single tones: synchrony and average discharge rate. Ph.D. Thesis, M.I.T., Cambridge, Mass.

KIANG, N.Y.S., 1975, Stimulus representation in the discharge patterns of auditory neurons, in The Nervous System, D.B. TOWER, ed., Raven Press, New York, Vol. 3, pp. 81-96.

KIANG, N.Y.S., MOXON, E.C., 1972, Physiological considerations in artificial stimulation of the inner ear. Ann. Otol. Rhinol. Laryngol. 81, pp. 714-730.

KIANG, N.Y.S., MOXON, E.C., 1974, Tails of tuning curves of auditory-nerve fibers. J.A.S.A. 55, pp. 620-630.

KIANG, N.Y.S., WATANABE, T., THOMAS, E.C., CLARK, L.F., 1965, Discharge Patterns of Single Fibers in the Cat's Auditory Nerve, MIT Press, Cambridge, Mass.

SCHROEDER, M.R., HALL, J.L., 1974, Model for mechanical to neural transduction in the auditory receptor. J.A.S.A. 55, pp. 1055-1060.
SIEBERT, W.M., GAMBARDELLA, G., 1968, Phenomenological model for a form of adaptation in primary auditory-nerve fibers. MIT, Res. Lab. Electron., Quart. Prog. Rep. 88, pp. 330-334.
SMITH, R.L., 1977, Short-term adaptation in single auditory-nerve fibers: some post-stimulatory effects. J. Neurophys. 40, pp. 1098-1102.
SMITH, R.L., ZWISLOCKI, J.J., 1975, Short-term adaptation and incremental responses of single auditory-nerve fibers. Biol. Cybernetics, 17, pp. 169-182.

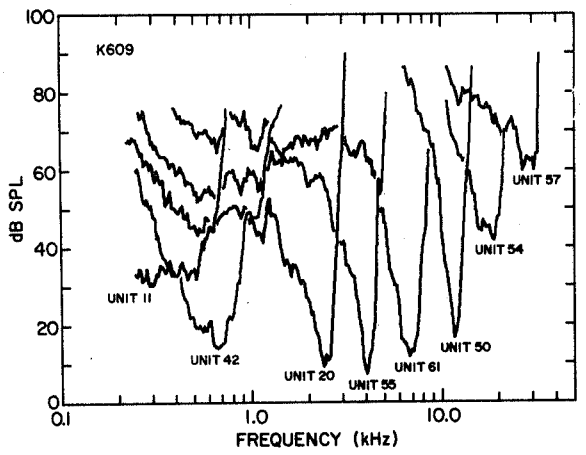


Fig. 1. (from KIANG,1975). Tuning curves for 8 fibers from one animal.
Courbes seuil-fréquence pour 8 fibres chez un même chat.

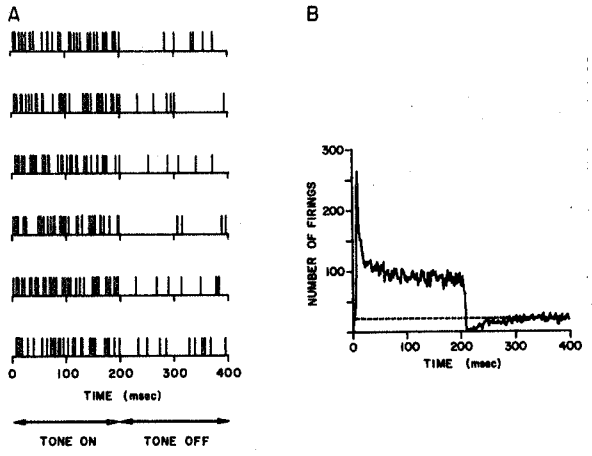


Fig. 2. A: Response of unit 17-17 to 6 presentations of a 200 msec tone burst followed by a 200 msec interval. Each spike is represented by a vertical bar. Tone-burst frequency: 2.08kHz. Level: 45 dB SPL. Threshold: 15 dB SPL. B: PST histogram computed from 300 presentations of the stimulus in A. Bin width: 2 msec. Vertical scale is number of firings in each bin. Dotted line is rate of spontaneous activity.
A: Réponse d'une fibre à 6 présentations d'un son pur de 200 msec suivi par un silence de 200 msec. Chaque PA est représenté par une barre verticale. B: Histogramme RTS calculé à partir de 300 présentations du stimulus de A.

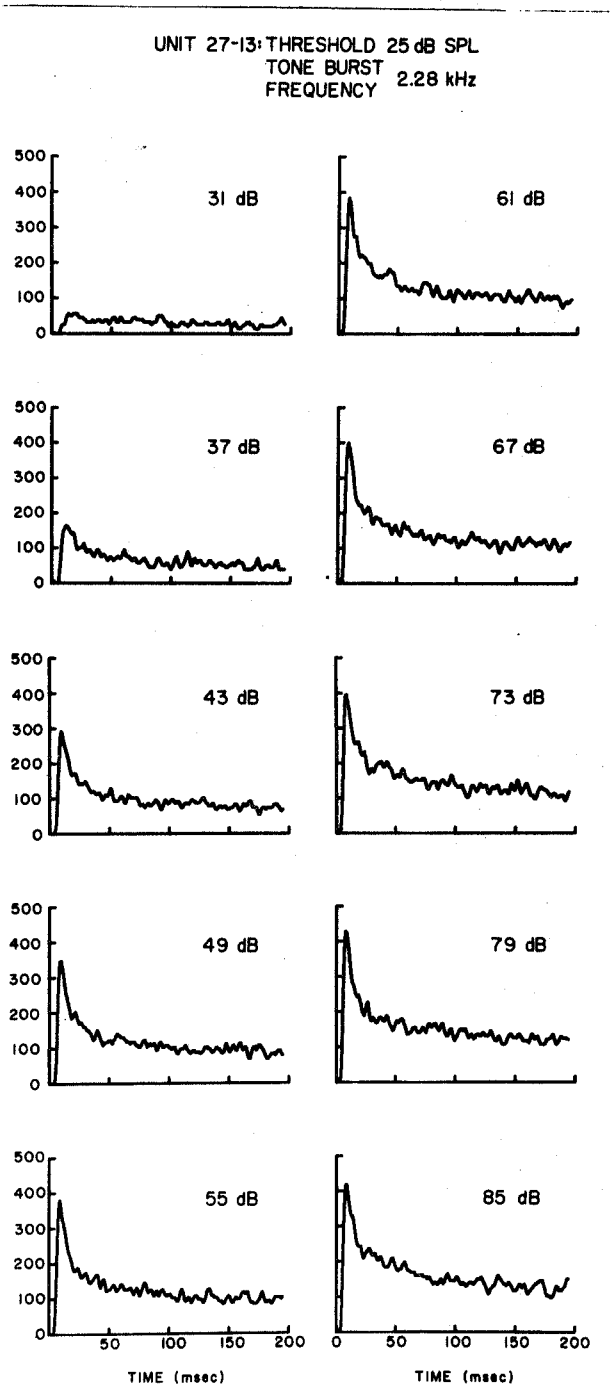
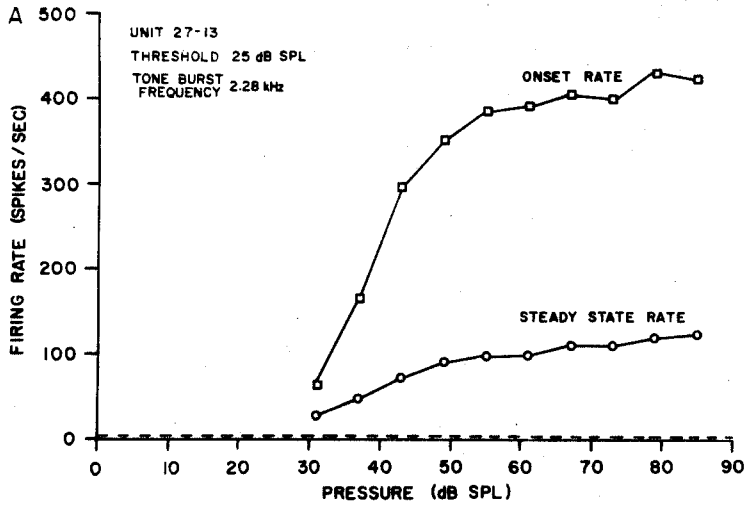


Fig. 3. PST histogram in response to a 200 msec tone burst at CF for 10 stimulus levels. Bin width: 1msec. Histograms were smoothed with a 6 msec Hamming window.
Histogrammes RTS en réponse à un son pur de 200 msec, pour 10 intensités du stimulus.



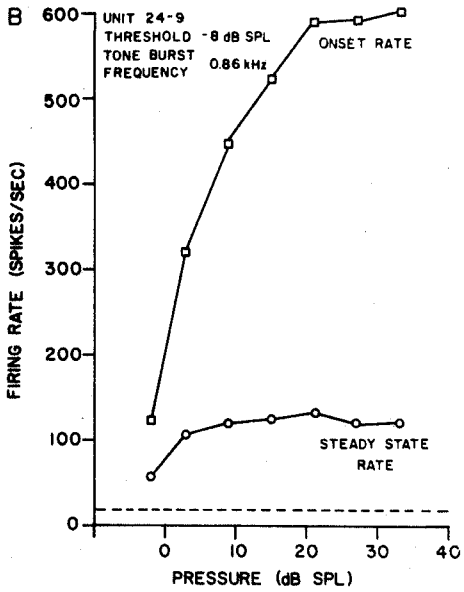


Fig. 4. A: Onset firing rate and steady-state rate vs level of tone burst at CF for unit of Fig. 3 The dotted line is the rate of spontaneous activity. B: Same as A for another unit.

A: Taux de déclenchement à l'attaque, et en régime stationnaire, en fonction de l'intensité du stimulus pour la fibre de la Fig. 3. L'echelle verticale est en PA/sec. La ligne pointillée représente le taux d'activité spontanée. B: Même chose pour une autre fibre.

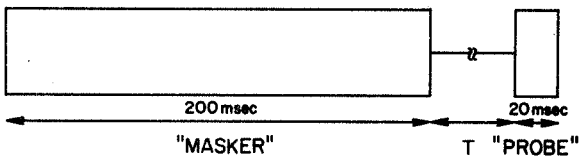


Fig. 5. Schematic diagram for two tone-burst paradigm. The probe was fixed at fiber CF, 20 dB above threshold, 20 msec in duration. The level and frequency of the masker, and the time delay T were varied.

Diagramme schématique du paradigme à deux sons purs. L'intervalle de silence T est une variable.

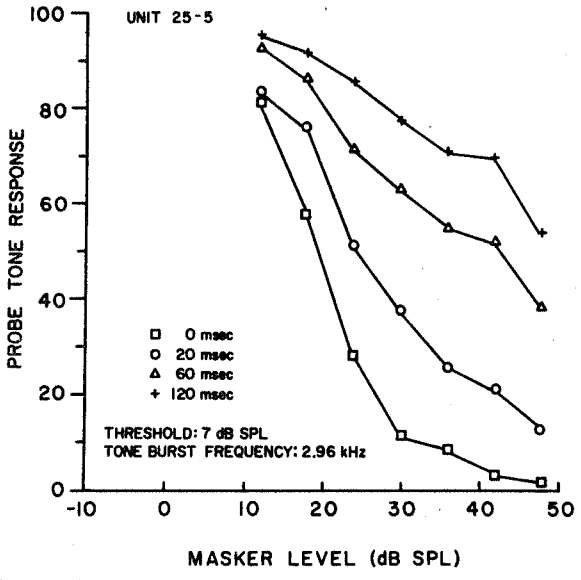


Fig. 6. Response to the probe-tone (in percent of the unmasked response) vs masker level, with time delay as parameter. The masker was at the fiber CF.

Réponse à la sonde (en pourcentage de la réponse en l'absence du masqueur) en fonction de l'intensité du masqueur, pour 4 valeurs de l'intervalle de silence. Le masqueur est à la FC.

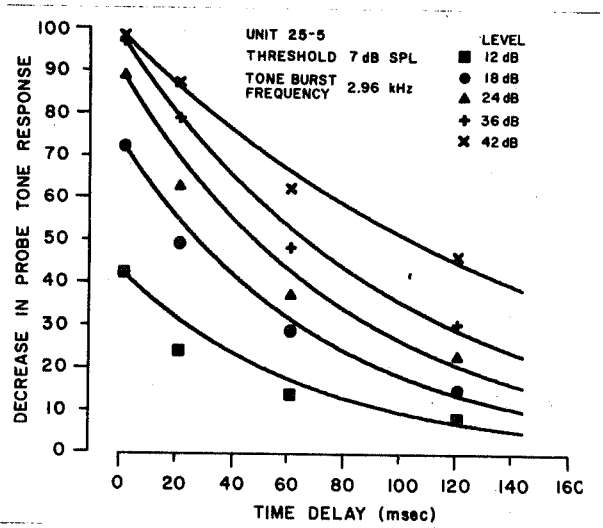


Fig. 7. Same data as in Fig. 6, plotted as the decrease in response to the probe-tone (in percent) vs time delay, with masker level as parameter. The 6 values of the masker level are given in dB relative to the threshold. The curves are exponential approximations to the temporal decay of masking for each masker level. Time constants of these exponentials are given below:

LEVEL	TIME CONSTANT
12 dB	67 msec
18 dB	72 msec
24 dB	82 msec
36 dB	99 msec
42 dB	153 msec

Mêmes points que sur la Fig. 6, reportés en fonction de la durée du silence pour 6 valeurs de l'intensité du masqueur (exprimées en dB au-dessus du seuil). L'ordonnée est la diminution en % de la réponse à la sonde, c'est-à-dire 100% moins l'ordonnée de la Fig. 6. Chacune des 6 courbes est une approximation exponentielle de la décroissance du masquage pour une intensité du masqueur. Pour les constantes de temps de ces exponentielles, voir le tableau ci-joint.

9èmes JOURNEES D'ETUDE SUR LA PAROLE

LANNION 31 mai - 2 juin 1978

PERCEPTION DES SEQUENCES DE CONTOURS PROSODIQUES DE PHRASES SYNTHETISEES

Philippe Martin

Experimental Phonetics Laboratory
University of Toronto

.**RESUME**

En se basant sur la congruence nécessaire entre les structures syntaxique et prosodique, il est possible de déterminer les ensembles de séquences de contours prosodiques acceptables ou inacceptables relativement à un énoncé donné. Pour obtenir des indications expérimentales quant à la validité des principes théoriques décrivant les relations syntaxe-intonation, on a soumis un groupe d'auditeurs à un test perceptif portant sur le caractère acceptable ou inacceptable de phrases synthétisées, porteuses de différentes séquences de contours mélodiques correspondant ou non aux séquences assignées par la théorie. Les résultats obtenus sont en concordance avec les prédictions théoriques, et leur interprétation mène à une caractérisation différentielle de certains syntagmes relativement à une mauvaise indication de la structure prosodique.

SUMMARY

Sequences of grammatical and non-grammatical prosodic contours can be predicted from a principle of congruence existing between syntactic and prosodic structures, the latter being always connected. In order to obtain experimental cues concerning the theoretical principles underlying the intonation-syntax relationship, a perceptual test has been conducted.

The degree of naturalness of 4 synthesized sentences bearing 23 different sequences of melodic contours has been evaluated by a panel of 7 listeners. These sequences were either linked to all possible prosodic structures not necesseraly congruent with the syntactic structure of the sentence, or completely irrelevant to the prosodic system of French. The results indicate that the theoretical assumptions are valid, and show certain properties of specific syntagms bearing wrongly located prosodic markers.

PERCEPTION DES SEQUENCES DE CONTOURS PROSODIQUES DE PHRASES SYNTHETISEES

Philippe Martin

Experimental Phonetics Laboratory
University of Toronto

INTRODUCTION

Des travaux récents sur la fonction phonosyntaxique de l'intonation permettent de déterminer un ensemble de séquences de contours prosodiques "grammaticaux" relativement à une structure syntaxique donnée (Martin, 1975). Partant, il est également possible de former des séquences de contours théoriquement inacceptables, soit parce qu'elles indiquent des structures prosodiques qui ne correspondent pas à la structure syntaxique de l'énoncé, soit parce que, ne dérivant pas de l'application des règles prosodiques propres au français, elles ne sont corrélatives d'aucune structure prosodique.

Pour tester la validité des principes théoriques décrivant la relation syntaxe-intonation, et qui portent essentiellement sur les propriétés de connexité et de congruence avec la hiérarchie syntaxique de la structure prosodique indiquée par les contours, on a soumis un groupe d'auditeurs à un test perceptif portant sur le degré d'"acceptabilité prosodique" de phrases synthétisées porteuses de différentes séquences de contours mélodiques correspondant ou non aux séquences assignées par la théorie. Ce test a donc trait non à la perception des contours prosodiques eux-mêmes, mais plutôt au fonctionnement de ces contours dans le cadre global de l'énoncé. Les résultats pourront donc s'interpréter relativement à la validité des mécanismes décrivant le rapport de la prosodie à la syntaxe.

INTONATION ET SYNTAXE

Dans sa fonction phonosyntaxique, l'intonation de la phrase peut être décrite par une séquence de contours mélodiques situés sur les seules syllabes accentuées des unités syntaxiques (et dont l'accent ne constitue pas une marque d'insistance). Ces contours apparaissent comme autant de marques indiquant une structure prosodique, c'est-à-dire un classement hiérarchique d'unités minimales (ou mots prosodiques) contenant un seul accent.

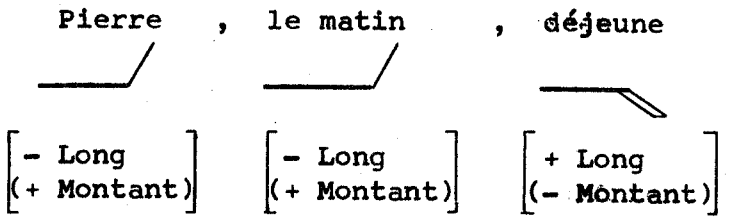
Alors que la structure syntaxique d'un énoncé n'est pas nécessairement connexe, la structure prosodique est, elle, en français, toujours connexe. Les unités prosodiques composant une unité de niveau supérieur dans la structure sont donc toujours contiguës dans la séquence.

Ainsi, la structure syntaxique de l'exemple

Pierre , le matin , déjeune

est non connexe, puisque les éléments du groupe (Pierre déjeune) sont séparés dans la chaîne de l'énoncé par l'unité (le matin), qui n'a de

lien syntaxique ni avec (Pierre) ni avec (déjeune). Par contre, la structure prosodique relative au même énoncé, et indiquée par une séquence de 3 contours décrits par les traits (- Long), (- Long) et (+ Long), est connexe, puisque les deux premières marques prosodiques - qui se manifestent généralement par des contours montants éventuellement suivis d'une pause - ne s'opposent pas entre eux.



La correspondance entre les mots prosodiques et les unités syntaxiques est décrite par une règle d'accentuation: une unité syntaxique minimale est accentuable et, partant, correspond à un mot prosodique, si elle ne contracte ni une relation de présupposition réciproque (solidarité) avec une autre unité (type Art. + Nom), ni plus d'une relation de présupposition unilatérale (sélection) avec une autre unité (type conjonction, pronoms relatifs, etc.) Dans ces derniers cas, l'unité syntaxique forme une seule unité accentuable soit avec l'unité dont elle est solidaire, soit avec l'unité qu'elle sélectionne à droite.

La correspondance entre les structures prosodique et syntaxique de l'énoncé est déterminée par un principe de congruence qui s'applique dans le cadre des contraintes géométriques imposées par la connexité de la structure prosodique. Composée de mots prosodiques qui peuvent être, de par la règle d'accentuabilité et par le jeu des désaccentuations possibles, plus grands que les unités syntaxiques minimales auxquels ils correspondent, la structure prosodique doit pouvoir inclure la structure syntaxique dans la mesure où la hiérarchie indiquée dans une structure ne peut contredire la hiérarchie des unités correspondantes indiquée dans l'autre structure.

S'il n'en n'est pas ainsi, du fait par exemple de la non connexité de la structure syntaxique à laquelle ne peut répondre la structure prosodique, il y aura dominance d'une des structures sur l'autre. Les marques de la structure dominée seront neutralisées dans leur participation à l'indication de l'autre structure (Martin, 1978).

TEST DE PERCEPTION

Le test de perception a porté sur le degré de naturel et d'acceptabilité prosodique de phrases synthétisées par concaténation de diphonèmes sur le système mis au point au C.N.E.T. par Emerard et Larreur (1976). Les stimuli comportaient 11 énoncés formés de 4 unités syntaxiques accentuables, présentant les différentes hiérarchies connexes possibles composées de 4 éléments.

A chacune de ces 11 structures, on a associé 23 séquences de contours mélodiques, parmi lesquelles 11 indiquaient les diverses structures prosodiques réalisables. Les 12 autres séquences correspondaient soit à des cas de désaccentuation des unités syntaxiques, soit à des réalisations

agrammaticales ne pouvant pas être générées par les règles propres au système du français.

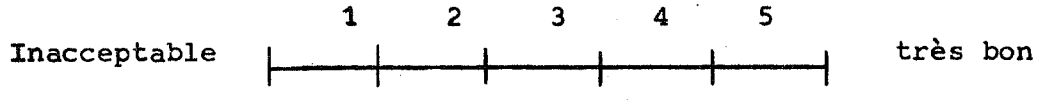
Chacun des contours prosodiques est associé à un symbole de commande du synthétiseur, qui définit en outre les variations de durée vocalique, d'intensité ainsi que d'éventuelles pauses selon le contexte phonétique propre à chaque unité synthétisée.

La mise en correspondance des 11 structures syntaxiques et des 23 séquences prosodiques est définie par les tableaux suivants:

STRUCTURES SYNTAXIQUES (indiquées par parenthésage)

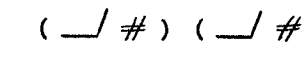
4A	(Metro) (boulot) Pernod) (dodo)
4B	((La soeur)(de Max)) ((mange)(les tartines))
4C	((Le frère)(de Marie)) (le matin) (déjeune)
4D	(Maurice)((le père)(de Julie)) (partira)
4E	(Paul) (le matin) ((avale)(son café))
4F	((Si Brigitte)(comme je le crains)(reviens)) (tu m'appelles)
4G	(((La fille)(du marin))(d'en face)) (est partie)
4H	((Si Pierre)((prend)(ses pilules))) (ça ira)
4J	(Le matin) ((Patrick)(souvent)(déjeune))
4K	(Le lundi) (((le bistrot)(d'en face))(est fermé))
4L	(Le matin) ((Juliette)((a pris)(sa valise)))

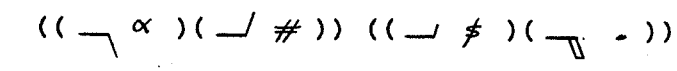
Le nombre de réalisations différentes étant relativement élevé (11 x 23 = 253), et chaque groupe de 23 séquences prosodiques devant être présenté deux fois afin d'établir un ancrage dans le jugement perceptif, on a limité le nombre de stimuli aux 23 variantes des énoncés 4B, 4C, 4D et 4E. 7 auditeurs ont noté le degré de naturel de ces phrases sur une échelle allant de 1 à 5 correspondant aux différentes qualités prosodiques perçues.

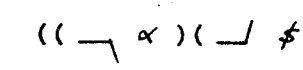


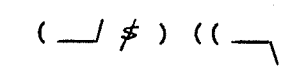
SEQUENCES PROSODIQUES

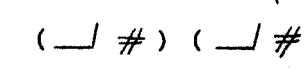
(les contours sont symbolisés par leur variation de fondamentale et associés aux symboles de commande du synthétiseur. Les structures prosodiques correspondantes éventuelles sont indiquées par parenthésage)

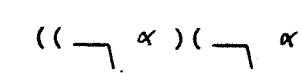
4a 

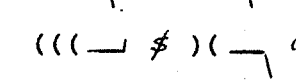
4b 

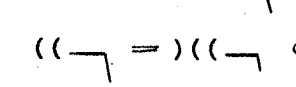
4c 

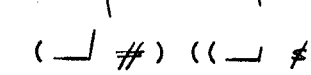
4d 

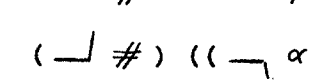
4e 

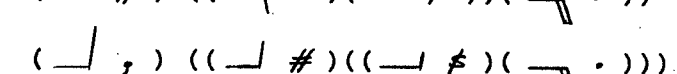
4f 

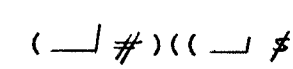
4g 

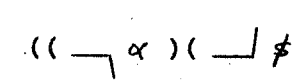
4h 

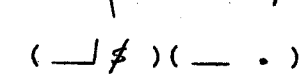
4j 

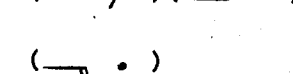
4k 

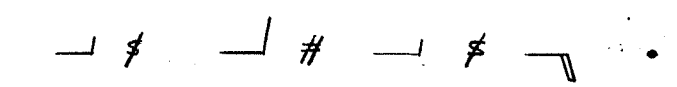
4l 

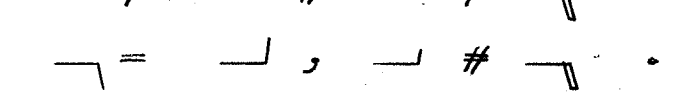
4m 

4n 

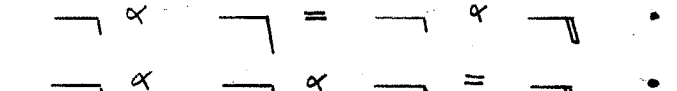
4p 

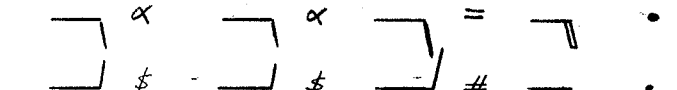
4r 

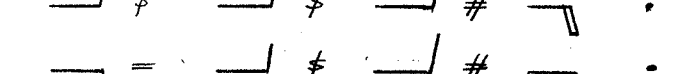
4s1 

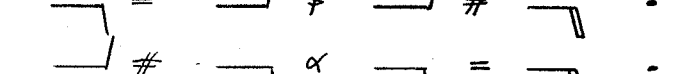
4s2 

4s3 

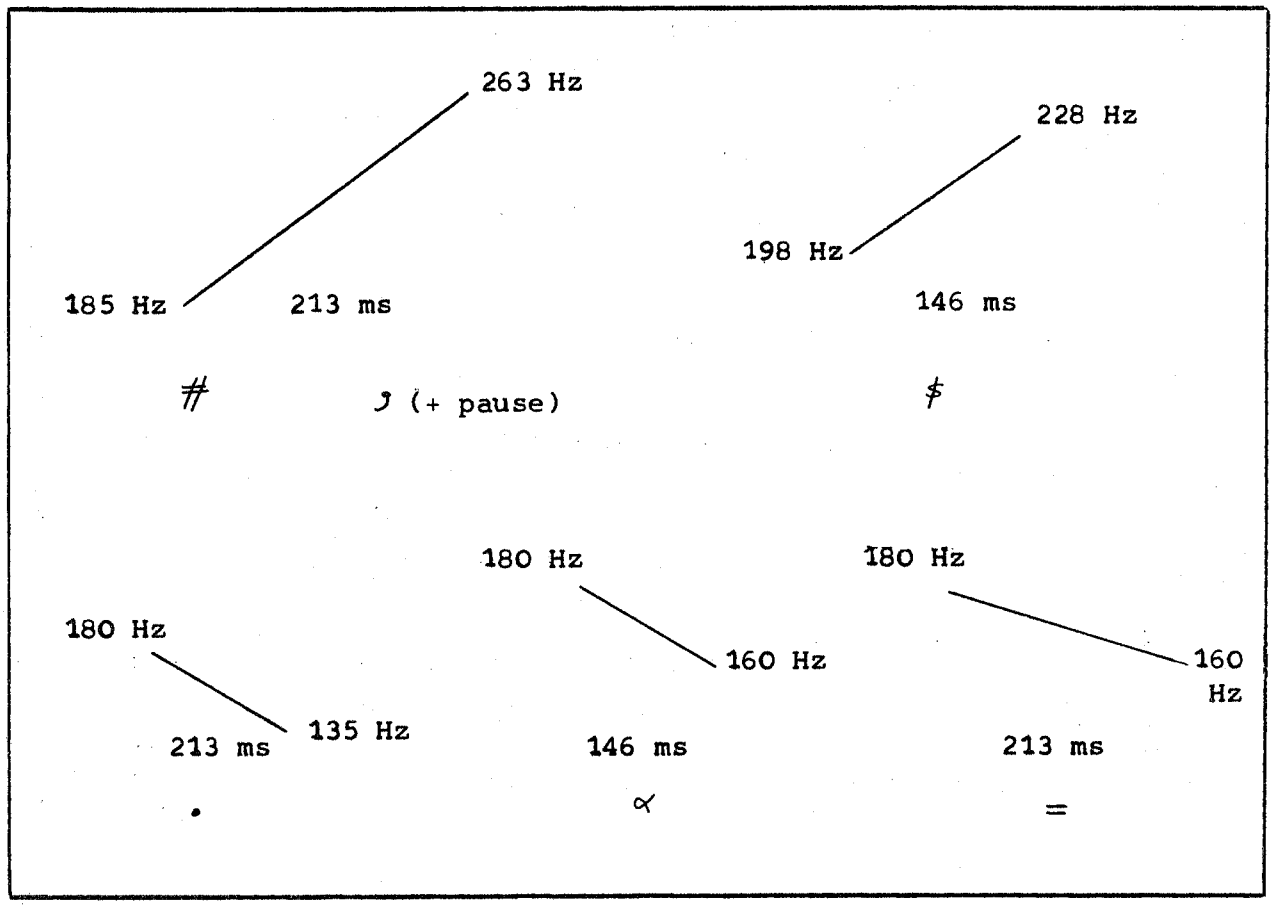
4t1 

4t2 

4t3 

4t4 

4v1 



RESULTATS

Pour chacune des réalisations jugées, on a calculé la moyenne des scores et effectué une analyse de variance selon la méthode de Scheffe. Dans ce but, les résultats ont été classés en 3 groupes: bons (moyenne supérieure à 3,50), moyens (score moyen compris entre 2,00 et 3,50) et mauvais (score moyen inférieur à 2,00).

Pour chacune des phrases, la séquence mélodique théorique, correspondant à la structure prosodique connexe congruente avec la structure syntaxique, a été trouvée significativement différente (au taux de 95%) du groupe des scores inférieurs à 2,00. Il en a été de même pour le groupe des jugements supérieurs à 3,50 opposés aux mauvaises réalisations, alors que la différence des groupes bons et mauvais opposés aux scores moyens est apparue non significative.

La composition des groupes obtenus se présente comme suit:

Structure 4B : La soeur de Max mange les tartines
(séquence prosodique théorique correspondante: 4b)

Bons : 4r; 4n; 4p; 4s3; 4c; 4e; 4s2; 4b
Moyens : 4t3; 4g; 4f; 4s1; 4j; 4t1; 4h; 4j; 4a
Mauvais : 4l; 4m; 4t4; 4k; 4v1; 4d

Structure 4C : Le frère de Marie, le matin, déjeune
(séquence prosodique théorique correspondante: 4c)

Bons : 4s2; 4b; 4m; 4c; 4p; 4l; 4s1;
Moyens : 4d; 4g; 4n; 4t2; 4f; 4r; 4t3; 4v1; 4h
Mauvais : 4k; 4s3; 4t4; 4j; 4e; 4a

Structure 4D : Maurice, le père de Julie, partira
(séquence prosodique théorique correspondante: 4d)

Bons : 4h; 4t3; 4k; 4j; 4t4; 4d
Moyens : 4r; 4f; 4g; 4v1; 4t2; 4l
Mauvais : 4l; 4a; 4c; 4s1; 4t1; 4n; 4m; 4b; 4s2; 4s3

Structure 4E : Paul, le matin, avale son café
(séquence prosodique théorique correspondante: 4e)

Bons : 4s1; 4m; 4s2; 4c; 4e; 4b
Moyens : 4l; 4a; 4d; 4p; 4f; 4n; 4t2; 4t3; 4g; 4j; 4r
Mauvais : 4k; 4s3; 4t1; 4h; 4v1; 4t4

L'examen de ces résultats mène aux quelques remarques suivantes:

- Conformément aux prédictions théoriques, les séquences correspondant aux structures prosodiques congruentes avec les structures syntaxiques ont été jugées parmi les meilleures (ceci valant également pour les structures à moins de 4 éléments obtenues par désaccentuation des unités syntaxiques);

- Les réalisations à un seul mot prosodique (type 4r) ont souvent reçu un score élevé, confirmant ainsi l'hypothèse voulant que l'absence de marque prosodique soit préférable à la présence d'un contour mal situé dans la structure;

- Les séquences prosodiques jugées inacceptables sont fréquemment corrélatives d'une non congruence entre les hiérarchies syntaxique et prosodique relativement au premier niveau des structures, mais cet effet ne semble pas jouer de manière symétrique dans la séquence. Ainsi, pour la phrase 4B, les auditeurs se sont montrés beaucoup plus sensibles à une mauvaise indication prosodique située au milieu du syntagme nominal sujet, alors que la présence de la même marque à l'intérieur du syntagme verbal a eu peu d'effet sur les scores. De même, les plus mauvais jugements de la phrase 4D sont dus à la présence d'une marque prosodique de premier niveau divisant le syntagme (le père de Marie), alors que la non indication des frontières de premier niveau situés après les unités (Maurice) et (le père de Marie) ont eu peu d'effet;

- Les réalisations non conformes au système du français (type 4s, 4t et 4v) ont été en général perçues comme peu naturelles. Les quelques cas contraires semblent indiquer une dominance des traits de pause et de durée vocalique dans l'indication d'une structure prosodique, les variations de fondamentale, non décodables selon les règles de la langue, ne jouant alors plus de rôle pour l'auditeur;

- L'hypothèse d'un effet d'emphase qu'entrainerait la présence, sur une unité déterminée, d'une marque prosodique incompatible avec la structure syntaxique ne semble pas se confirmer ici; les rares effets perçus étant dus à la présence de pauses dans des séquences jugées acceptables.

CONCLUSIONS

La perception de stimuli verbaux dépendant de nombreux facteurs difficilement contrôlables, et compte tenu du faible nombre de sujets impliqués, l'interprétation des résultats obtenus ici ne peut constituer qu'un indice de validité plutôt qu'une démonstration de la théorie ainsi testée. Il reste que, dans l'ensemble, les propriétés de congruence et de connexité qui, avec les mécanismes d'indication de la structure prosodique, définissent les options théoriques retenues, ne sont pas prises en défaut.

L'éventail plus large que prévu des réalisations jugées acceptables pourrait s'expliquer par les variations relativement faibles de fréquence fondamentale utilisées dans le système de synthèse, variations fréquemment insuffisantes pour éviter la neutralisation des contours relativement à une structure syntaxique dominante. Un montage expérimental plus élaboré devrait permettre d'éclaircir ce point, et d'aborder l'étude de structures plus complexes, ainsi que celle des valeurs phonétiques à attribuer aux contrastes existant entre les marques prosodiques.

REFERENCES

EMERARD, F. , 1977, Synthèse par diphones et traitement de la prosodie, Thèse de 3ème cycle, Université de Grenoble II.

EMERARD, F. et LARREUR, D. , 1976, Synthèse par diphones, Recherches/ Acoustique, C.N.E.T., Lannion, III, pp. 293-314.

MARTIN, Ph. , 1975, Analyse phonologique de la phrase française, Linguistics, 146, pp. 35-67.

MARTIN, Ph. , 1978, Résumé d'une théorie de l'intonation, Bull. de l'Inst. de Phonétique de Grenoble, avril 1978.

9èmes JOURNEES D'ETUDE SUR LA PAROLE

LANNION 31 mai - 2 juin 1978

PERCEPTION DE LA QUANTITE VOCALIQUE ET CONSONANTIQUE EN ITALIEN, PAR ANALYSE-SYNTHESE DE LA VOIX

Enrico Vivalda — C.S.E.L.T.
Pier Marco Bertinetto — Université de Turin

.RESUME

On sait que l'italien présente un méchanisme de compensation dans les durées respectives d'une voyelle accentuée et de la consonne qui suit. Par conséquent, des mots tels que *papa* vs *pappa* peuvent être représentés de la façon suivante: /'pa:pa/ vs /'pap:a/. Un problème débattu dans la phonologie italienne est celui qui concerne la détermination de l'élément pertinent dans cette relation.

Pour repondre à cette question, nous avons envisagé l'expérience suivante. Un locuteur a prononcé une courte liste de mots du type /'CV:CV/ vs /'CVC:V/. Les matériaux originels ont étés analysés au moyen d'une technique de prédiction linéaire, et ensuite resynthétisés avec des variations systématiques dans la durée de la voyelle ou de la consonne, ou bien de ces deux segments à la fois. Les stimuli ainsi obtenus ont été soumis, en ordre casuel, à une équipe de sujets pour un test de discrimination perceptive. Les réponses nous permettent d'affirmer que l'élement pertinent est à chercher dans la durée consonantique. Nous avons aussi déterminé un seuil d'inversion dans les jugements catégoriels.

ON THE PERCEPTION OF VOWEL AND CONSONANT LENGTH IN ITALIAN. A RESEARCH BY ANALYSIS-SYNTHESIS TECHNIQUES

Enrico Vivalda — C.S.E.L.T.
Pier Marco Bertinetto — University of Turin

.SUMMARY

It is generally assumed that the Italian language exhibits a kind of balancing mechanism among the respective lengths of a stressed vowel and the following consonant. Such words as *papa* vs. *pappa* might therefore be transcribed as /'pa:pa/ vs. /'pap:a/. A largely debated problem in Italian phonology is deciding whether the relevant feature, for the categorization of such linguistic stimuli, is to be found in vowel or consonant length: the remaining element should of course be regarded as automatically conditioned.

To provide an answer to this dilemma, we have conceived the following experience. A short list of words of the type /'CV:CV/ vs. /'CVC:V/ has been pronounced by a single speaker. The original materials have been processed by linear prediction techniques, and subsequently resynthesized with a systematic variation of the duration of either the stressed vowel or the following consonant, or both. The stimuli, which have been suitably randomized, have been submitted for a forced-choice discrimination to a group of listeners. Their responses enable us to state that consonant duration is the relevant factor of this compensatory relation. We have also been able to individuate the shifting-point of these categorizing judgements.

INTRODUCTION

On a depuis longtemps observé que l'italien présente un phénomène très marqué de compensation dans les durées vocaliques et consonantiques. Cela peut s'exprimer de la façon suivante: la durée d'une voyelle accentuée est affectée par la durée de la consonne intervocalique qui suit. Ce phénomène, qu'on a souvent appelé "isochronisme syllabique" (bien que cette désignation soit acceptable ou non selon le cadre théorique qu'on veut adopter), n'est d'ailleurs pas une charactéristique exclusive de l'italien: on le retrouve tel quel, par exemple, dans les langues scandinaves. Le problème se pose maintenant de savoir quel est l'élément qui gouverne le relation de durée entre la voyelle et la consonne suivante. Deux solutions se présentent tout de suite à l'esprit:

(a) on peut considérer la voyelle comme l'élément pertinent de la relation, et la consonne comme l'élément automatiquement déterminé;

(b) on peut enverser cette condition de détermination, en regardant la consonne comme l'unite douée de valeur phonologique en ce qui concerne l'attribution du trait "quantité".

Comme on le sait, pour les langues scandinaves on a généralement adopté la solution (a); ce qui est d'ailleurs en accord avec les résultats instrumentaux obtenus par Elert [1965] pour le suédois. En ce qui concerne l'italien, on a traditionnellement choisi la solution (b) [Muljačic 1972]. Saltarelli [1970] a quand même avancé récemment l'hypothèse que la solution (a) soit adoptable pour l'italien aussi, surtout pour des raisons d'économie dans la description du système phonologique. Mais Bertinetto [à paraître] a demontré, à travers l'ana-

lyse spectrographique des productions verbales de quatre sujets, que la solution (b) est à préférer. En particulier, cette étude revèle que, pour des paires de mots du type ['CV:CV] vs ['CVC:V] prononcés en contexte, la dispersion des consonnes autour de la durée moyenne est bien plus limitée que la dispersion des voyelles accentuées. Cela signifie sans doute que la durée des consonnes est un facteur beaucoup plus stable et efficace pour la décodification, pendant que la durée des voyelles accentuées (surtout de celles en syllabe ouverte) se charactérise par un marge de variation bien plus large.

Il nous semble très intéressant, maintenant, de vérifier quelles sont les réactions perceptives des sujets natives à propos des questions que nous avons exposées. Plus exactement, on peut se demander quel serait l'effet d'une manipulation systématique des rapports de durée entre les voyelles et les consonnes dans les mots disyllabiques dont on a parlé. C'est dans ce but que nous avons envisagé les deux expériences qui font l'object de ce travail.

Remarque. Nous laissons ici de côté la question très débattue dans la phonologie italienne, qui concerne le statut de /C:/. Bien que nous pensons qu'il s'agisse du redoublement d'un même phonème, nous allons le considérer dans la suite de ce travail, pour des raisons de simplicité, comme un seul phonème long.

Dans le même souci de simplicité, on va ici adopter la convention d'indiquer par [V:] et [V] les voyelles accentuées, ainsi que par [C] et [C:] les consonnes intervocaliques des mots considérés dans notre étude.

DESCRIPTION DE LA TECHNIQUE INSTRUMENTALE

Pour réaliser nos expériences, on doit avoir la possibilité de modifier les durées de certains segments, appartenant à des mots préalablement enregistrès, sans entâmer sensiblement les charactéristiques spectrales du son. Une tecnique efficace nous est offerte par les méthodes d'analyse-synthèse de la voix. Dans notre recherche, nous avons employé un programme intéractif, développé chez le CSELT [Bertinetto, Scagliola, Vivalda 1977], qui permet d'effectuer l'analyse, la synthèse et la segmentation du signal. L'analyse, basée sur la technique de la prédiction linéaire, nous donne, tous les 8 msec, les paramètres essentiels pour décrire le procès phonatoire, c'est-à-dire: 10 coéfficients de réflexion k_i, un paramètre binaire de reconnaissance "son voisé / non-voisé", la valeur de la fréquence fondamentale, la valeur de l'intensité sonore, et la valeur de la cadence d'extraction des paramètres. Tous ces données sont stockées dans la mémoire de l'élaborateur. La synthèse est obtenue en envoyant les paramètres à un synthétiseur (simulé en software), qui consiste dans un filtre en échelle.

Si l'on utilise la partie du programme conçue pour la segmentation, on peut localizer, au moyen d'un vidéo display, la portion du signal sur laquelle on veut opérer. Les modifications des paramètres sont obtenues en envoyant les instructions appropriées par la consolle. Ainsi, en variant la cadence d'extraction des paramètres, nous avons modifié la durée des voyelles et des consonnes sans toucher les transitions du spectre.

EXPERIENCE I

Un des auteurs de ce travail (PMB, qui parle une variété nordoccidentale de l'italien) a enregistré sur bande magnétique, dans une chambre sourde, les cinq suivants paires de mots: *papa* (='pape') vs *pappa* (='popote'), *Cana* (='Cana')

vs *canna* (='canne'), *sbafo* (='je mange avidement') vs *sbaffo* (='trace, tache'), *cacio* (='fromage') vs *caccio* (='je chasse'), *pala* (='pelle; pale; retable') vs *palla* (='balle'). Tous les dix mots ici énumérés peuvent se présenter dans le vocabulaire italien courant, bien que la probabilité d'occurrence soit différente selon les cas; en particulier, *Cana* est possible seulement dans l'expréssion biblique *les noces de Cana, sbafo* appartient à la langue parlée, *cacio* est usé surtout dans les régions centrales de l'Italie. Comme on le voit, la voyelle accentuée est toujours /a/ dans nos exemples; la consonne intervocalique, au contraire, a été choisie de façon qu'un certain nombre des principales classes consonantiques seraient représentées: occlusives, nasales, fricatives, affriquées, liquides. Il est important de noter ici que la prononce de *cacio* , pour les locuteurs italiens du nord, est ['katʃo]; c'est-à-dire, il s'agit d'une véritable affriquée.

Les mots, synthétises par la téchnique qu'on a décrite auparavant, ont été modifiés de la façon suivante. On a d'abord calculé pour chaque paire de mots la différence de durée entre [V] et [V:] d'une part, [C] et [C:] de l'autre. Ensuite, on a trouvé pour chaque paire de mots quatre pas de variation (chacun de même longueur), tels qui rendaient une [V] aussi longue qu'une [V:], ou bien une [C] aussi longue qu'une [C:]. Enfin, on a synthétisé, pour chaque mot, huit nouveaux exemplaires, en suivant la procédure ici bas décrite. Soit d_v la différence de durée qui sépare [V] de [V:], d_c celle qui sépare [C] de [C:]:

(i) pour les mots du type ['CV:CV] on a successivement (a) raccourci [V:] de 1/4 d_v, 2/4 d_v, 3/4 d_v et 4/4 d_v; (b) allongé [C] de 1/4 d_c, 2/4 d_c, 3/4 d_c et 4/4 d_c (ainsi, par exemple, une [V:] diminuée de 4/4 d_v posséde évidemment la même durée que la [V] correspondante dans l'autre membre du paire en question);

(ii) pour les mots du type ['CVC:V] on a adopté la procédure inverse, avec allongement de la voyelle accentuée et raccourcissement de la consonne intervocalique.

En total, on a ainsi obtenu seize échantillons différents, qui unis aux deux originels (également synthétisés) nous donnent dix-huit stimuli pour chaque paire de mots: c'est-à-dire, 90 stimuli pour l'ensemble des cinq paires. Il est à remarquer ici que, dans le cas des voyelles, les variations ont été calculées de manière que les transitions restaient intactes, afin que la qualité du son n'était pas affectée par les modifications apportées. Il faut noter aussi que le locuteur a eu le soin de prononcer tous les deux mots de chaque paire au même niveau mélodique, de façon que l'identification de chaque stimulus n'était pas influencée par des facteurs prosodiques autres que la relation de durée entre voyelle et consonne suivante. Du reste, si cela était le cas, on aurait pu aisément corriger le déroulement de la ligne mélodique au moment de la synthèse. Une dernière considération s'impose à propos de la méthode employée dans la modification des durées. Comme le but de notre étude était d'évaluer les effets perceptifs, nous avons décidé de privilégier le rapport entre les différentes durées intrinsèques des voyelles et des consonnes dans chaque paire de mots, plutot que les valeurs absolus des pas de variation en tant que tels. Ainsi, quoique d_v et d_c varient selon les divers cas, la variation pourcentuelle est toujours la même.

Les 90 stimuli ont été transférés, en ordre casuel, sur une bande magnétique, et présentés en chambre acoustiquement traitée à 24 sujets (étudiants chez l'Université de Turin), pour un test de discrimination perceptive. Les sujets devaient indiquer, sur un papier préalablement préparé, s'ils avaient antendu un mot de l'un ou de l'autre type, le choix étant obligé. Avant le début du

test, l'expérimenteur a soigneusement commenté chaque mot, de façon que tous les sujets informateurs se convaincraient de la légitimité de les prononcer dans les contextes appropriés. Les sujets ont été de même invités à ne pas se faire influencer par la nature des réponses qu'ils avaient déja données à chaque instant, le choix devant etre absolument concentré sur le particulier stimulus qu'ils venaient d'entendre. Dans ce but, on a insisté que, à cause des variations apportées aux échantillons originels, les sommes des jugements de l'un et de l'autre type ne devaient vraisemblablement pas être égales. Il faut quand même ajouter que la véritable nature des variations a été bien sûr tenue cachée aux informateurs, afin de ne pas influencer leur réponses.

Les résultats de ce test sont reproduits en fig. 1, a ÷ d. L'ordonnée des quatre graphiques montre le pourcentage des jugements; qui sont on faveur du type ['CV:CV] en 1a et 1b, et en faveur du type ['CVC:V] en 1c et 1d. L'abscisse montre les quatre pas de variation adoptés dans chaque cas, correspondants au progressif raccourcissement de [V:] en 1a et de [C:] en 1d, et au progressif allongement de [C] en 1b et de [V] en 1c. L'extension de ces pas de variation a été conventionellement égalisée dans tous les graphiques, bien que (comme nous l'avons noté plus haut) en réalité leur dimensions étaient différentes selon l'amplitude de d_v et d_c dans chaque paire de mots.

Comme on le voit très bien en 1a et 1c, les modifications apportées à la durée de [V:] et [V] n'entrainent presque aucune conséquence sur la décodification des stimuli. En effet, quoique les changements du rapport de durée entre la voyelle et la consonne suivante soient très marqués, ces échantillons ont été presque toujours identifiés comme faisant partie de la même classe que le mot originel (c'est-à-dire, le mot a partir duquel on a apporté la variation). En ce qui concerne [C] et [C:], on peut affirmer le contraire: les variations dans la durée de la consonne entraînent très rapidement une complète inversion des jugements. Ainsi, en 1b on passe de ['CV:CV] à ['CVC:V], et la situation opposée se présente en 1d.

Cela nous permet de dégager une première conclusion: la durée de la consonne intervocalique est de loin le facteur le plus important pour la discrimination, tandis que la voyelle joue un rôle beaucoup plus marginal. Tout cela résulte aussi dans la fig. 2, où nous avons rapporté les pourcentages moyennes des jugements relatifs aux consonnes intervocaliques. Plus exactement, le côté gauche de la figure (courbe *a*) presente les données moyennes du graphique 1b; tandis qu'en lisant la meme fig. 2 de droite à gauche (courbe *b*) on trouve les résultats moyens du graphique 1d. On voit ici que l'allure des courbes est celle qu'on observe typiquement dans la circonstance d'un complet renversement des réponses perceptives. Mais le fait qui nous intéresse surtout de vérifier dans ce cas est l'éventuelle présence d'une dissymmétrie dans le déroulement de ces deux types de catégorisation. En effet, le point de croisement des courbes est légèrement déplacé vers le côté gauche de la figure; et cela veut dire que la modification de la durée de [C] gagne plus rapidement le seuil d'inversion en comparaison avec la semblable modification de [C:]. Evidemment on a affaire ici à des facteurs autres que le simple rapport de durée entre voyelle et consonne. Il est à supposer que la différente tension articulatoire de [C] et [C:] soit la cause principale de cette frappante dissymmétrie.

La dernière conclusion qu'on peut tirer des résultats obtenus jusqu'ici concerne la nature intrinsèque du son consonantique impliqué dans les différents paires de mots. On voit de façon très claire en figure 1b et 1d que la dispersion des données est parfois très remarquable. Les consonnes qui semblent s'éloigner le plus de la moyenne sont l'occlusive et la fricative, respecti

vement représentées dans notre corpus par /p/ et /f/. Il faut noter en particulier que la conduite de ces deux consonnes est très cohérente dans les deux graphiques 1b et 1d: un allongement relativement petit de /p/ cause une dramatique inversion dans les jugements perceptifs, tandis que le raccourcissement de /p:/ doit être bien considérable pour obtenir un résultat semblable. Avec /f/ on observe la situation opposée. Cela pourrait nous amener à penser que le seuil inverseur de [C] en [C:] (et vice versa) se place différemment dans les deux cas: plus voisin de [C] avec l'occlusive, plus voisin de [C:] avec la fricative. D'ailleurs, la longueur de la voyelle et de la consonne dans ces deux paires de mots compris dans notre corpus est très semblable, comme on dégage du tableau suivant:

Tableau I

['CV:CV]		['CVC:V]	
[a:] = msec 280;	[p] = msec 176	[a] = msec 168;	[p:] = msec 352
[a:] = msec 288;	[f] = msec 184	[a] = msec 176;	[f:] = msec 328

Il faut toutefois se garder de tirer des conséquences trop nettes de ces données. Il ne faut pas oublier que nos stimuli ont été construits à partir d'une simple exécution individuelle de chaque mot. Donc, il se peut que le différent comportement des diverses consonnes soit ici déterminé, en partie au moins, par des facteurs de nature tout à fait contingente. Avant d'élaborer sur ce point une véritable théorie, il faudra bien sûr faire de nouvelles expériences; par example, en construant les échantillons originels de façon telle, que les durées des segments vocaliques et consonantiques soient calculées en rapport aux valeurs moyennes dégagées des productions de plusieurs locuteurs. Il nous suffira ici d'avoir démontré que le problème existe.

EXPERIENCE II

Dans les stimuli qui on fait l'objet de la première expérience, nous avons modifié la durée d'un seul segment à la fois: celle de la voyelle, ou bien celle de la consonne. Cela aboutit sans doute à une grossière altération de la structure prosodique du mot originel. Nous avons donc envisagé une deuxième expérience, dans laquelle nous avons varié conjointement la longueur de ces deux segments. Le paire qu'on a choisi pour ce test est *papa* vs *pappa* , le même qui était déja présent dans notre corpus. Dans le tableau suivant nous avons dressé la liste des seize nouveaux stimuli, construits par la même technique instrumentale illustrée plus haut, auxquels on doit ajouter les deux stimuli originels. Comme les échantillons de départ sont les mêmes qu'on avait utilisés dans le test précédant, d_v et d_c préserve ici les mêmes valeurs qu'auparavant. Le tableau indique aussi, pour chaque stimulus, le rapport C/V, qui devient dans ce cas l'objet principal de la recherche. Comme on le voit, le stimulus 5 présente évidemment le même rapport C/V que le stimulus 11, puisque les durées des segments sous examen sont parfaitement égalisées. Une observation tout à fait pareille peut être faite à propos des stimuli 1 et 15.

Les sujets qui ont pris part à ce deuxième test sont les mêmes qui avaient déjà participé à la première expérience. Les résultats son montrés dans la fig. 3. Ici, nous avons rapportés en ordonnée les pourcentages des réponses en faveur de *papa* (voir la courbe *a*) et en faveur de *pappa* (voir la courbe *b*). Sur l'abscisse on peut lire les valeurs du rapport C/V pour les divers stimuli, qui sont à leur tour identifiables au moyen du numéro d'ordre (voir tableau II).

Tableau II

PAPA

STIMULUS	V:	C	RAPPORT C/V
1	V:	C	0.62
2	V: - 1/4 d_v	C + 1/4 d_c	0.87
3	V: - 2/4 d_v	C + 2/4 d_c	1.18
4	V: - 3/4 d_v	C + 3/4 d_c	1.57
5	V: - d_v	C + d_c	2.09
6	V: - 1/4 d_v	C + 3/4 d_c	1.22
7	V: - 3/4 d_v	C + 1/4 d_c	1.12
8	V: - 1/4 d_v	C + 2/4 d_c	0.92
9	V: - 2/4 d_v	C + 1/4 d_c	0.98

PAPPA

STIMULUS	V	C:	RAPPORT C/V
11	V	C:	2.09
12	V + 1/4 d_v	C: - 1/4 d_c	1.57
13	V + 2/4 d_v	C: - 2/4 d_c	1.18
14	V + 3/4 d_v	C: - 3/4 d_c	0.87
15	V + d_v	C: - d_c	0.62
16	V + 1/4 d_v	C: - 3/4 d_c	1.12
17	V + 3/4 d_v	C: - 1/4 d_c	1.22
18	V + 1/4 d_v	C: - 2/4 d_c	1.34
19	V + 2/4 d_v	C: - 1/4 d_c	1.37

On voit très bien que les courbes respectivement dessinées par les stimuli 1-9 et 11-19 se croisent dans un point qui se trouve tout près de la valeur 1.10 du rapport C/V. De plus, on peut noter que, au delà de la zone délimitée par les valeurs 0.87 et 1.22 de ce rapport, on observe une remarquable unanimité dans la catégorisation des stimuli. En effet, tous les jugements compris dans les deux bandes latérales du graphique, au delà de la zone critique, sont significatifs au niveau 0.01 selon le test du χ^2 (c'est-à-dire, on n'a qu'une chance sur cent d'obtenir ces mêmes résultats d'une façon purement casuelle).

Il reste à expliquer la brusque infléxion dessinée par la courbe *a* entre les stimuli 2, 8, 9, 7. Il n'est pas aisé d'indiquer la solution de cet énigme. On peut remarquer toutefois que, bien que le rapport C/V soit presque identique dans les échantillons 8 et 9, le premier présente une longueur consonantique décidemment plus marquée en comparaison avec la longuer vocalique. Il est donc parfaitement compréhensible que les jugements des sujets soient poussés vers la catégorie de *pappa* plutôt que de *papa*. Le contraire vaut évidemment pour le stimulus 9. D'ailleurs, les résultats qu'on a obtenus ici sont tout à fait confirmés par la fig. 4; où nous avons rapporté (en utilisant la même procédure employée pour la fig. 3) les données observées dans la première expérience pour les mots appartenents au paire *papa* vs *pappa*. Il est à observer que dans cette figure nous avons dressé deux type de courbes: des courbes pleines, qui représentent la conjonction de tous les timuli, e des courbes pointillées, qui

unissent seulement les points relatifs aux échantillons obtenus par variation de la durée consonantique. Il va de soi que les timuli 1 et 11 (c'est-à-dire, les deux stimuli de départ) sont les mêmes soit dans la fig. 3 que dans la fig. 4. Deux choses sont particulièrement frappantes ici: (i) le point de croisement des courbes se trouve à nouveau tout près de la valeur 1.10 du rapport C/V; (ii) la courbe de gauche revèle, entre les stimuli 4, 7, 5 et 8, une infléxion tout à fait semblable à celle que nous venons de commenter à propos de la fig. 3. Ce qui est intéressant est que la structure durative du mot 7 de la fig. 4 est très voisine de celle du mot 8 de la fig. 3 (voir le tableau II): en effet, dans le stimulus 7 de la fig. 4 la voyelle accentuée possède la valeur de départ [V:], tandis que la consonne intervocalique possède la valeur "[C] + 2/4 d_c ". Il faudra évidemment approfondir ce point dans une nouvelle expérience. Mais pour l'instant nous pouvons affirmer avec confiance que, pour chaque consonne donnée (ici, /p/) la valeur du rapport C/V semble être gouvernée par des lois bien rigoureuses dans les mots du type que nous avons examinés dans cette étude. En effect, pour obtenir le résultat perceptif désiré, il est nécéssaire de se tenir au delà d'une zone critique très limitée. D'autre part, on ne doit pas négliger non plus la longueur absolue de la consonne comme facteur d'identification: la discussion que nous venons de faire à propos des infléxions de la courbe *a* dans les figures 3 et 4 le démontre clairement. Et cela confirme une fois de plus que la consonne est le véritable trait pertinent dans les oppositions de "quantité" en italien.

CONCLUSION

Nous croyons pouvoir affirmer que ce travail nous a apporté des connaissances nouvelles sur la structure prosodique de disyllabes italiens du type ['CV:CV] vs ['CVC:V]. Les conclusions les plus importantes peuvent être resumées de la façon suivante.

(1) La durée de la consonne est de loin la variable la plus décisive dans la discrimination perceptive des mots impliqués dans une opposition de "quantité" en italien. Cet effet est en outre renforcé par la différente tension articulatoire respectivement associée à [C] et [C:]. Donc, la thèse traditionnelle (défendue aussi par Bertinetto [à paraître]) résulte confirmée, contre l'opinion de Saltarelli, d'ailleurs très suggestive du point de vue théorique.

(2) Pour chaque type de consonne intervocalique, il semble qu'on peut définir avec une remarquable exactitude des seuils critiques dans le rapport C/V, bien que la valeur absolue de la durée de la consonne intervocalique joue parfois un rôle autonome dans la décodification.

Nous nous proposons d'approfondir, dans des recherches qui sont en cours à présent, les aspects inhérents en particulier au point (2) ci-dessus.

REFERENCES

Bertinetto, P.M. [à paraitre], "La quantità vocalica in italiano: verifica spettrografica", à paraitre dans les Actes du XIV Congrès Int. de Ling. et Phil. Romane (Naples 1974).

Bertinetto, P.M., Scagliola, C., Vivalda, E. [1977], "Interactive text-to-speech system for the determination of Italian prosodic rules" International Congress on Acoustic(Madrid 1977).

Elert, C. [1965], *Phonologic studies of quantity in Swedish,* Uppsala.

Muljačic, Z. [1972], *Fonologia della lingua italiana,* Bologna.

Saltarelli, M. [1970], *A phonology of Italian in a generative grammar,* The Hague.

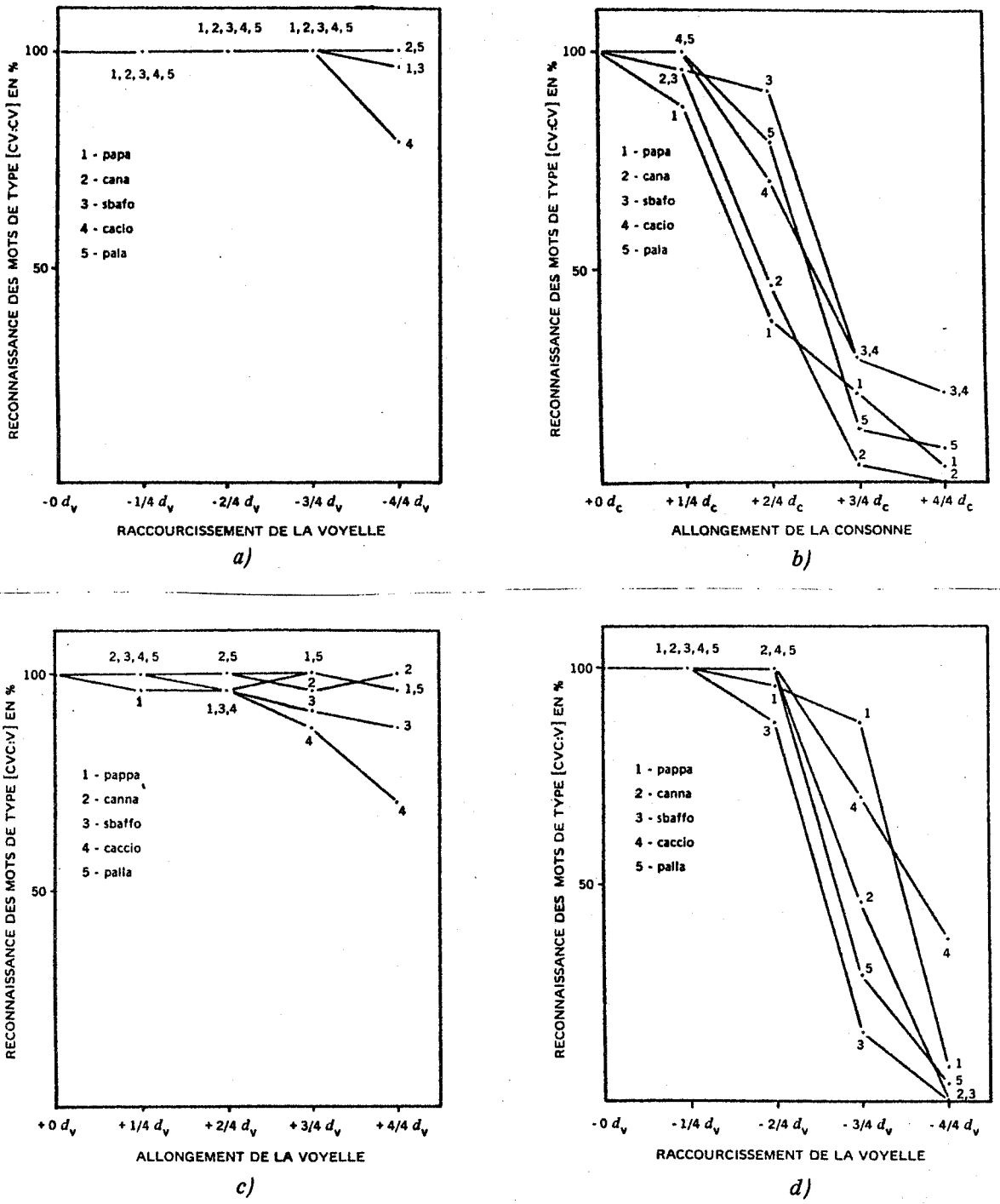


Fig. 1 - Pourcentages des jugements de reconnaissance obtenues dans le test I, en fonction de la variation proportionnelle des durées des voyelles accentuées et des consonnes intervocaliques. Les quatre figures montrent respectivement:

a) raccourcissement de [V:] dans les mots de type ['CV:CV]

b) allongement de [C] intervocalique dans les mots de type ['CV:CV]

c) allongement de [V] accentuée dans les mots de type ['CVC:V]

d) raccourcissement de [C:] dans les mots de type ['CVC:V].

Dans le cas de a) et b), les pourcentages sont en faveur du type ['CV:CV]; dans c) et d) elles sont en faveur du type ['CVC:V]. En outre: d_v = différence de durée entre [V:] et [V] dans chaque paire de mots; d_c = différence de durée entre [C:] et [C] .

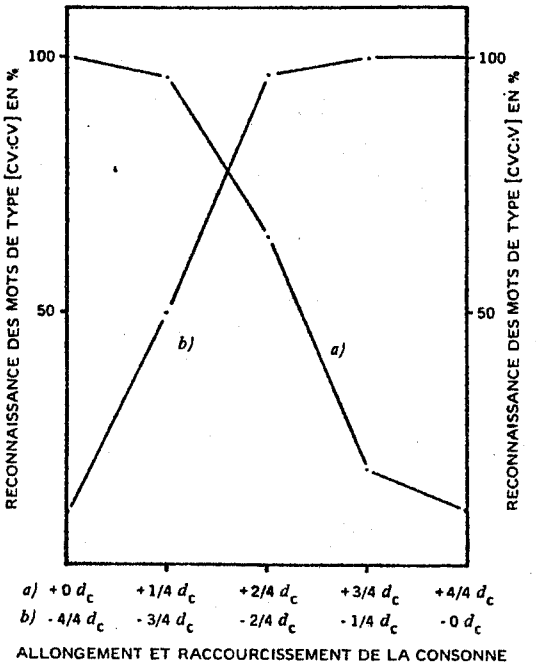


Fig. 2 - Pourcentages moyennes des jugements relatifs aux cinq paires de stimuli, en fonction des variations apportées à la durée de la consonne intervocalique (courbes moyennes dégagées des figures 1b et 1d). La courbe de gauche se refère aux mots de type ['CV:CV], celle de droite aux mots de type ['CVC:C]. L'abscisse indique respectivement, pour la courbe a), l'allongement de [C], et pour le courbe b), le raccourcissement de [C:].

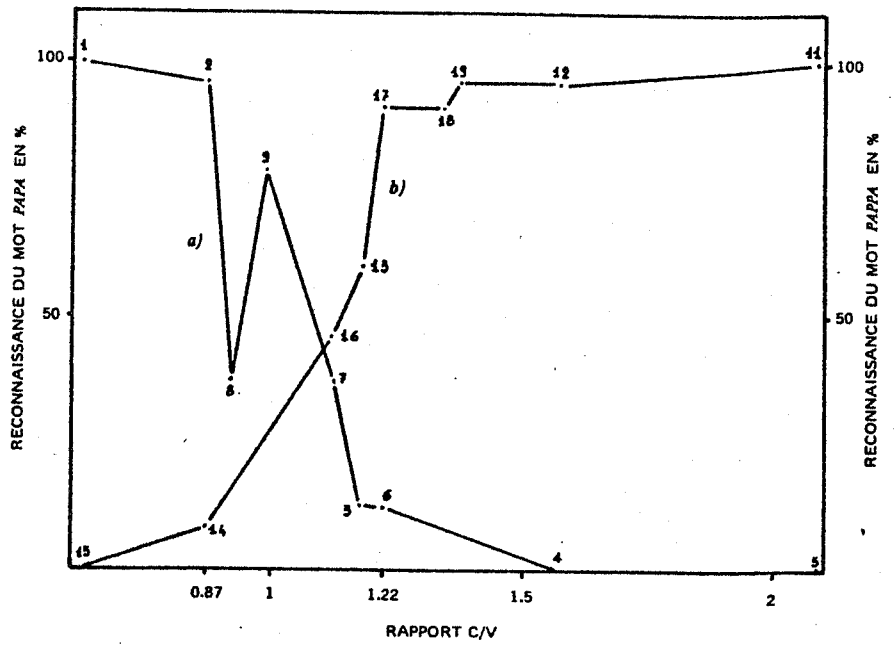


Fig. 3 - Pourcentages de reconnaissance des mots «papa» (courbe a) et «pappa» (courbe b), en fonction du rapport C/V (durée de la consonne intervocalique divisée par la durée de la voyelle accentuée). Ces données se referènt aux stimuli utilisés dans l'expérience II

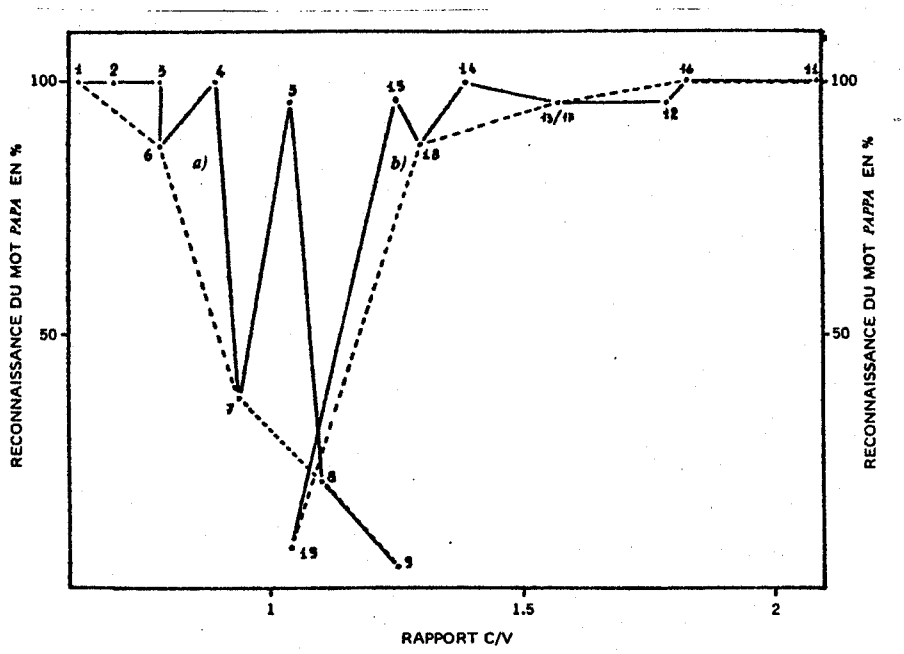


Fig. 4 - Pourcentages de reconnaissance des mots «papa» (courbe a) et «pappa» (courbe b), en fonction du rapport C/V. Ces données se refèrent aux stimuli utilisés dans l'expérience I. Les courbes pointillées unissent seulement les échantillons obtenus par modification de la durée consonantique

9èmes JOURNEES D'ETUDE SUR LA PAROLE

LANNION 31 mai - 2 juin 1978

ETUDE COMPAREE DE LA SENSIBILITE DIFFERENTIELLE A LA DUREE AVEC UN SON PUR ET AVEC UNE VOYELLE

BOVET Pierre Laboratoire de Psychologie (L.A. CNRS 181)

ROSSI Mario Institut de Phonétique, Aix-en-Provence (L.A. CNRS 261)

.RESUME

Ce travail concerne l'étude du seuil différentiel pour les sons purs et les voyelles, pour des durées brèves de 200 ms. On utilise une méthode d'autorégulation appliquée à une situation de choix forcé entre 2 éventualités. A chaque essai un seul stimulus est présenté et le sujet doit se prononcer sur la durée de ce stimulus en choisissant l'un des réponses : court ou long. Cette méthode est caractérisée par le fait que la durée effective du stimulus présente à chaque essai est déterminée en fonction des performances du sujet aux essais précédents. Le système informatique utilisé dans l'expérience est adapté au contrôle de processus expérimentaux en phonétique et en psychophysique.

Aucune différence de sensibilité n'a pu être mise en évidence entre sons purs et voyelles. Le seuil trouvé pour ces deux types de stimuli est de l'ordre de 10 %. Nos résultats sont en conformité avec ceux de Fujisaki 1973 et de Nooteboom (1972) ; par ailleurs ils révèlent l'unicité du mécanisme mis en jeu dans la perception différentielle de la durée. La perception de la durée apparaît comme un phénomène central indépendant de détecteurs spécialisés.

COMPARATIVE STUDY OF THE DIFFERENTIAL SENSITIVITY TO DURATION IN PURE SOUNDS AND VOWELS

BOVET Pierre — Laboratoire de Psychologie (L.A. CNRS 181)

ROSSI Mario — Institut de Phonétique, Aix-en-Provence (L.A.CNRS 261)

SUMMARY

This study is concerned with the differential threshold for pure sounds and vowels for short durations of 200 ms. The method employed is one of auto-regulation applied to a situation of forced choice between two alternatives. At each trial the subject is presented with a single stimulus and is required to judge it either short or long. A characteristic feature of this method is that the actual duration of the stimulus presented at each trial is determined in function of the subject's performance on previous trials. The computer system used in the experiment is adapted for the control of experimental procedures in phonetics and psychophysics.

The results showed no significant difference between the sensitivity for pure sounds and that for vowels. The threshold for both types of stimuli was around 10 %. Our results confirm those of Fujisaki (1973) and Nooteboom (1972); they also point to a single mechanism being involved in the differential perception of duration. The perception of duration consequently appears to be a central mechanism independant of specialised detectors.

INTRODUCTION

Ce travail concerne l'étude du seuil différentiel pour les sons purs et les voyelles. La recherche porte sur des durées brèves, de l'ordre de 200 millisecondes. En effet, ce sont les durées de cet ordre de grandeur qui sont les plus fréquentes dans la parole naturelle pour la réalisation des voyelles sous l'accent, notamment à la fin de la phrase.

Une étude comparée de la sensibilité différentielle pour les sons purs et les sons de parole, pourrait permettre d'apporter quelque lumière sur le problème de savoir si la perception de la durée relève ou non du même mécanisme selon que le système auditif traite des sons dont la nature et la fonction sont aussi différentes que celles des sons sinusoïdaux et des sons complexes de la parole.

En ce qui concerne les seuils différentiels de fréquence et d'intensité dans le domaine auditif, on relève d'importantes différences entre les sons sinusoïdaux et les sons de parole en faveur des premiers. Cette constatation a conduit ZWICKER (1962) à supposer l'existence pour l'audition des sons purs, d'un mécanisme auditif spécifique d'une très haute sensibilité pour l'intégration des changements de fréquence des sons sinusoïdaux. Inversement, il est permis de faire l'hypothèse que l'organisme humain est mieux adapté à la perception des sons de parole qu'à celle des sons purs, conformément aux travaux sur la perception catégorielle (LIBERMAN et al. 1961).

Il reste à voir si les recherches consacrées à la durée confirment ou non l'une ou l'autre de ces hypothèses.

En ce qui concerne les valeurs du seuil différentiel de durée obtenues par nos prédecesseurs, on relève des divergences importantes entre les résultats, sans que l'on puisse noter de différences systématiques entre son pur et son de parole. Ces divergences doivent être imputées essentiellement à la méthode et éventuellement à l'intensité des stimulus (tous les auteurs n'indiquent pas l'intensité des stimulus qu'ils utilisent...). Il ne semble pas que les divergences puissent être imputées à la fréquence comme l'ont montré SMALL et CAMPBELL (1962)et RHUM et al. (1966).

METHODE EXPERIMENTALE UTILISEE DANS NOTRE RECHERCHE

On trouvera cette méthode décrite et illustrée dans BOVET (1977 a et b). Il s'agit d'une méthode d'autorégulation appliquée à une situation de choix forcé entre 2 éventualités. A chaque essai un seul stimulus est présenté : stimulus court (s = 1) ou stimulus long (s = 2). Et le sujet doit se prononcer sur la durée de ce stimulus en choisissant respectivement l'une des 2 réponses (r = 1 ou 2).

Cette méthode est caractérisée par le fait que la valeur effective (durée physique) du stimulus présenté à chaque essai est déterminée en fonction des performances du sujet aux essais précédents. Cette durée du stimulus s_n présentée à un essai n, durée que nous notons $//s_n//$ dépend bien entendu principalement de la catégorie du stimulus ($s_n = 1$ ou $s_n = 2$) : le choix de cette catégorie étant aléatoire et indépendant des performances du sujet aux essais précédents. Mais la valeur $//s_n//$ est également déterminée par une variable d, dépendante des performances aux essais précédents, qui correspond à la difficulté de l'identification du stimulus : d a pour valeur, à chaque essai n, la différence entre les durées des 2 stimulus qui peuvent être présentés (court

ou long), on a :

$$d_n = // s_n^2 // - // s_n^1 //$$

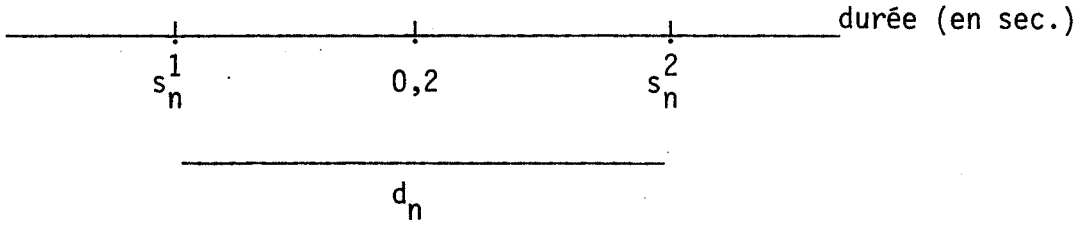
Par ailleurs les valeurs des 2 stimulus qui peuvent être présentées à un essai sont toujours symétriquement disposées par rapport à une valeur centrale constante, valeur autour de laquelle nous cherchons à déterminer la sensibilité différentielle : 0,2 s.

Ainsi, en définitive, la durée (en secondes) du stimulus présenté à l'essai n est déterminée par la règle suivante :

$$// s_n // = 0{,}2 - \frac{d_n}{2} \text{ , si } s_n = 1$$

$$// s_n // = 0{,}2 + \frac{d_n}{2} \text{ , si } s_n = 2$$

règle qui peut être représentée ainsi :



Le modèle qui préside aux variations de d est fondé sur le comportement d'un sujet idéal dont le seuil différentiel serait constant et égal à une certaine valeur δ . Dans ce cas hypothétique, on obtiendrait, par définition du seuil différentiel, une matrice de confusion comportant 3 fois plus de réponses correctes que de réponses fausses. La règle d'expérimentation est telle que d_n, dans ce cas, tendrait à se stabiliser autour de δ :

Soit, pour n = 2,3,... :

$$d_n = d_{n-1} - a \text{ , si } r_{n-1} = s_{n-1} \qquad \text{(réponse juste)}$$

$$d_n = d_{n-1} + 3\,a \text{ , si } r_{n-1} = s_{n-1} \qquad \text{(réponse fausse)}$$

a est une constante dont la valeur est un paramètre de l'expérience. La régulation est optimisée par le choix d'un pas trois fois plus petit dans le cas de réponses justes (trois fois plus fréquent) que dans le cas de réponses fausses.

Nous préciserons plus loin (§§ IV et V) les valeurs de d_1 et de a retenues pour chacune de nos deux expériences.

On voit que, tant que l'écart inter-stimulus dn est supérieur au seuil différentiel, il tend à diminuer, car la proportion de réponses justes est supérieure à 75 %, tandis que tant que cet écart est inférieur au seuil, il tend à augmenter.

La figure 2 représente l'évolution de d_n dans le cas d'un sujet idéal.

Une fois stabilisée, la quantité d_n peut ainsi être utilisée pour estimer

le seuil différentiel dans le cas où celui-ci est inconnu.

En fait, nous n'utilisons pas $\underline{d}$ comme estimation du seuil mais un autre indice $\underline{dh}$ - fonction des 4 derniers essais - qui ne fluctue pas mais reste stable et égal au seuil dans le cas d'un sujet idéal.

On trouvera justifié le choix de notre indice dh dans BOVET (1977 a) :

$$dh_n = \max (d_j / r_i \neq s_i \; ; \; j = n-3, n-2, n-1, n) + \frac{3}{2} a$$

En fin de compte, on prendra comme valeur de sensibilité différentielle, la valeur minimale atteinte par $\underline{dh}$.

Au demeurant, il nous faut encore préciser que dans notre expérience le sujet est constamment tenu au courant de sa performance, par un score qui est une fonction linéaire décroissante de $\underline{dh}$. Ce score est affiché, à chaque essai, sur un écran de visualisation placé en face du sujet.

Plus précisément, le sujet a connaissance, sur l'écran :

1°) du caractère juste ou faux de la réponse qu'il vient de donner,
2°) du score instantané qu'il a obtenu, fonction de dh_n ,
3°) du score maximal qu'il a atteint depuis le début de la séance expérimentale.

DISPOSITIF EXPERIMENTAL

MATERIEL

Nos expériences sont pilotées par un ordinateur de laboratoire (SEMS T 1600) couplé à 2 clés electroniques (Grason stadler 1287B), à un clavier de 3 boutons tout-ou-rien et à un écran de visualisation vidéo. Le clavier et l'écran, ainsi qu'un haut-parleur sont placés en face du sujet dans une chambre anéchoïque.

L'ensemble de production des stimulus sonores comporte, en amont des clés electroniques, respectivement un générateur de basses fréquences et un synthétiseur de parole paramétrique ; en aval il est composé d'un atténuateur, d'un mixeur et d'un amplificateur.

Le système informatique utilisé dans l'expérience, avec tous ses interfaces, (contrôle des sources, horloge de cadencement, distribution vidéo, clavier de réponses), est adapté au contrôle de processus expérimentaux en phonétique et en psychophysique (TESTON, 1975).

LOGICIEL

Les programmes régissant nos expériences ont pu être rédigés en langage évolué (FORTRAN IV) grâce à l'utilisation d'une bibliothèque de sous-programmes spécifiquement destinée à l'expérimentation en psychologie (DUQUENNE, CNRS, Univ. R. Descartes).

CARACTERISTIQUES DES SIGNAUX

Le son pur est un signal sinusoïdal de 150 Hz, à 70 dB au-dessus du seuil (2×10^{-4} dyne/cm2) mesuré par la chaîne Brüel et Kjoer étalonnée.

Le son de parole est la voyelle [a] produite par le synthétiseur à partir d'un train d'impulsions en dents de scie approximant l'onde glottale ; les caractéristiques de la voyelle sont les suivantes :

$$Fo = 150 \text{ Hz}, F_1 = 850 \text{ Hz}, F_2 = 1200 \text{ Hz}, F_3 = 2400 \text{ Hz}$$

L'intensité physique est la même que pour le son pur.

Tous les signaux (son pur ou voyelle) quelle que soit leur durée sont modulés au début et à la fin, selon une pente constante (passage de 0 dB à 70 dB - ou inversement - en 25 ms) afin d'éviter tout effet de "clic" aux bornes des stimuli.

On notera enfin la caractéristique suivante du montage expérimental : le générateur de basses fréquences et le synthétiseur fonctionnent en permanence, et les instants d'apparition (par fermeture de l'une ou l'autre clé electronique) de chacun des signaux successifs (stimulus) dépendent du sujet. C'est ce dernier,en effet, qui provoque l'apparition des signaux en appuyant sur un bouton central.

On comprend que, de la sorte, la segmentation des signaux sur une période soit parfaitement aléatoire, et par là, rendue indépendante de leur durée. En d'autres termes, aucune information sur la durée - objet de notre étude - n'est susceptible d'être transmise par l'intermédiaire d'éventuels indices liés à la segmentation.

EXPERIENCE 1

PLAN

4 sujets, hommes adultes habitués aux expériences de laboratoire, ont subi chacun successivement et à quelques jours d'intervalle, 4 séances expérimentales. Dans cette première expérience, il n'est présenté lors d'une séance qu'un seul type de signaux : soit des sons purs (séances 2 et 3), soit des voyelles (séances 1 et 4). On a $d_1 = 0{,}04$ s (= 20 %), et $a = 4 \times 10^{-3}$s. Lors de chacune des séances, la succession des stimuli 1 et des stimuli 2 (cf. supra § II) est complètement aléatoire (équiprobabilité).

Cette expérience présente d'ailleurs la propriété de comporter un nombre quelconque d'essais dans chaque séance expérimentale ; le nombre d'essais dépend du sujet : en effet, c'est le sujet qui décide de la fin d'une séance expérimentale lorsqu'il estime qu'il ne pourra plus améliorer son score maximal.

RESULTATS

Le tableau 1 présente l'évolution globale de l'indice de sensibilité différentielle. Il n'en ressort, après analyse, aucun effet significatif dû au type de signaux utilisés. La tendance d'un effet d'appentissage (accroissement moyen de la performance de la première à la dernière séance) n'est elle-même pas significative.

On doit donc conclure que notre première expérience ne permet pas de mettre en évidence une quelconque différence de sensibilité différentielle aux durées entre un son pur et une voyelle.

Par ailleurs, si on l'interprète comme un seuil différentiel relatif, la valeur moyenne observée de l'indice (= 7 %) est plus fine que les valeurs du seuil calculées par nos prédécesseurs (sauf RUHM et al, 1966 et FUJISAKI, 1973).

Ce résultat peut s'expliquer par le fait que notre mesure de la sensibilité différentielle est fondée sur une valeur minimale correspondant à une performance limite plutôt qu'à une moyenne.

EXPERIENCE 2

PLAN

5 adultes, 4 hommes et 1 femme, subissent cette fois une seule séance expérimentale ; les paramètres sont fixés comme suit :

d_1 = 0,03 s , ce qui correspond à un écart inter-stimulus relatif initial de 15 % valeur considérée comme supra-liminaire pour tous les sujets retenus ;

$a = 0{,}4 \times 10^{-3}$s, valeur très petite permettant cependant d'atteindre, dans le cas où toutes les réponses sont justes, la zone du seuil différentiel attendu (7,5 %) en un nombre raisonnable d'essais (< 40).

L'autorégulation est, comme précédemment, fondée sur 2 catégories de stimulus : inférieurs à 0,2 s (s = 1), ou supérieurs à 0,2 s (s = 2). Mais ici, chacun de ces 2 types de stimulus peut être communiqué au sujet soit sous la forme du son pur, soit sous la forme de la voyelle, au cours de la même séance expérimentale. De la sorte on a affaire à 4 événements stimulants : pur court, pur long, voyelle courte, voyelle longue.

La suite des stimulus présentés à un sujet est formée à partir de ces 4 événements équifréquents. Une séance comporte dans cette deuxième expérience un nombre constant d'essais ($512 = 2 \times 4^4$) dont les stimuli sont organisés selon le principe des mots circulaires équilibrés (cf. BOVET, 1975 ; et DURUP, 1967). Cette technique, qui préserve une parfaite imprévisibilité des événements stimulants pour chacun des sujets, permet également un contrôle extrêmement précis des éventuels effets séquentiels.

RESULTATS

L'autorégulation fonctionnant tout à fait indépendamment de la "nature" des stimuli (son pur ou voyelle), nous ne pouvons plus ici utiliser l'indice dh pour comparer la sensibilité différentielle à la durée (1).

Nous présentons dans le tableau II le nombre de réponses correctes par sujet et par type de signal : cette présentation permet de comparer immédiatement la sensibilité différentielle aux sons purs et aux voyelles. Les tests statistiques présentés dans le tableau montrent bien qu'ici encore aucune différence de sensibilité n'a pu être mise en évidence entre le son pur et la voyelle.

On notera (tableau III) que ce résultat global se conserve si l'on ne considère plus que les stimulus précédés par un stimulus de même nature (sons purs précédés par un son pur, ou voyelles précédées par une voyelle). On ne peut donc évoquer, pour expliquer cette identité de sensibilité différentielle, l'effet simple de contagion d'un type de signal sur l'autre.

---

(1) Au demeurant les valeurs minimales de dh obtenues ici sont du même ordre de grandeur que celles trouvées dans l'expérience 1. Soit, respectivement pour chaque sujet : 8,1 % ; 14,3 % ; 5,9 % ; 5,1 % ; 8,5 %.

Si l'on s'intéresse maintenant (tableau IV) non plus au nombre de réponses correctes ou erronées mais au nombre de réponses court (r = 1) ou long (r = 2), on observe une différence systématique de traitement suivant la nature des stimulus : sons purs ou voyelles. Les voyelles sont systématiquement surestimées par rapport au son pur.

Cette surestimation peut être interprétée comme le résultat du poids spectral de la voyelle qui véhicule une information paradigmatique très riche qui interfère avec l'axe syntagmatique.

C'est le même phénomène qui lie étroitement intensité et durée (MUNSON 1947, ZWICKER 1973).

Il est vain d'ailleurs d'ouvrir une polémique pour savoir si c'est le poids spectral ou l'intensité qui est l'élément essentiel, car ROSSI (1971) a montré que l'intensité spécifique des voyelles était une conséquence de la répartition spectrale de l'énergie. Il semble évident que la surestimation en question est due à l'intensité spécifique de la voyelle, intensité qui est elle-même la conséquence d'une plus grande richesse spectrale.

CONCLUSION

Les résultats auxquels nous aboutissons ne permettent pas d'étendre à la perception de la durée l'hypothèse émise par SWICKER en ce qui concerne la fréquence et l'amplitude : les sons purs ne sont pas traités de façon privilégiée pour ce qui est de leur durée. Ces résultats confirment ceux de FUJISAKI qui ne trouve pas de différence significative dans la sensibilité différentielle de la durée avec des stimulus différents allant des sons purs aux sons de parole. Ceci revient à dire que le seuil différentiel de durée pour les voyelles est relativement fin : inférieur à 10 % lorsqu'il est mesuré avec notre méthode ou même celle de FUJISAKI. Cette constatation permet de souligner l'importance du rôle de la durée dans le langage. Des expériences de production et de perception de la parole menées par NOOTEBOOM (1972) à partir de stimulus identiques, montrent que les variations systématiques de durées observées en production, sont également pertinentes en perception. La finesse de discrimination perceptive de la durée est du même ordre que celle de la régulation articulatoire. Pour les durées comprises entre 80 et 150 ms elle se situe autour de 8 %. De plus, les sujets qui font preuve de la précision la plus grande sur le plan articulatoire sont également ceux dont la sensibilité différentielle est la plus fine. Production et perception dans la parole sont bien en relation étroite. Et si ce fait n'implique pas nécessairement le recours à la théorie motrice, du moins on peut dire que ces 2 modes sont étroitement adaptés.

On ne peut être que frappé par le fait que nous retrouvons ici des valeurs de seuil identiques à celles de NOOTEBOOM.

Mais plus que l'identité numérique de nos résultats avec ceux de NOOTEBOOM et de FUJISAKI, notre recherche révèle l'unicité du mécanisme mis en jeu dans la perception différentielle de la durée quelle que soit la nature du stimulus. La perception de la durée apparaît comme un phénomène central indépendant de détecteurs spécialisés.

REFERENCES

BOVET, P., Génération automatique de mots circulaires et équilibrés, Mathématiques et sciences humaines, 49, 1975, pp. 29-41.
a) Mesure de la sensibilité différentielle par une procédure autorégulée : I. Principe de la méthode, Travaux de l'Institut de Phonétique d'Aix, 4, 1977, pp. 7-27.
b) Mesure de la sensibilité différentielle par une procédure autorégulée : II. Programmation d'une expérience, Travaux de l'Institut de Phonétique d'Aix, 4, 1977, pp. 31-41.

CREELMAN, C.D., Human discrimination of auditory duration, J.A.S.A., 34, 5, 1962, pp. 583-593.

DURUP, H., Graphes et plans d'expériences temporels, mots circulaires et plans toriques, Mathématiques et Sciences Humaines, 18, 1967, pp. 1-31.

FUJISAKI, H., NAKAMURA, K. and IMOTO, T., Auditory perception of duration of speech and non speech stimuli, Symposium on auditory analysis and perception of speech,Leningrad, 1973, pp. 1-27.

HUGGINS, A.W.F., Just noticeable differences in segment duration in speech, J.A.S.A., 51, 1972, pp. 1270-1278.
On the perception of temporal phenomena in speech, J.A.S.A., 51, 1972, pp. 1279-1290.

KLATT, D.H. and COOPER, W.E., Perception of segment duration in sentence contexts, Structure and process in speech perception, ed. Cohen et Nooteboom, Springer, Berlin, 1975, pp. 69-89.

KOZHEVNIKOV, V.A. and CHISTOVICH, L.A., Speech : Articulation and perception, Clearinghouse, Washington, 1966.

LIBERMAN, A.M., HARRIS, K.S., EIMAS, P., LISKER, L. and BASTIAN, J., An effect of learning in speech perception : the discrimination of durations of silence with and without phonemic significance, Language and speech, 4, 1961, pp. 175-195.

MUNSON, W.A., The growth of auditory sensation, J.A.S.A., 19, 4, 1947, pp. 584-591.

NOOTEBOOM, S.G., Production and perception of vowel duration, thèse, Ed. de l'Université, Eindhoven, 1972.

ROSSI, M., L'intensité spécifique des voyelles, Phonetica, 24, 3, 1971, pp. 129-161.
Le seuil différentiel de durée, Papers in linguistics and phonetics to the memory of Pierre Delattre, La Haye, 1972, pp. 435-450.

RUHM, H.B., MENEKE, E.O., MILBURN, B., COOPER, W.A. and ROSE, D.E., Differential sensitivity to duration of acoustic stimuli, Journal of speech and hearing research, 9, 1966, pp. 371-384.

SMALL, A.M., and CAMPBELL,R.A., Temporal differential sensitivity for auditory stimuli, American journal of psychology, 75, 1962, pp. 401-410.

TESTON, B., Description d'un système informatique adapté au contrôle de processus expérimentaux et à l'instrumentation, Cahiers de psychologie 18, 1975, pp. 127-147.

ZWICKER, E., Direct comparisons between sensations produced by frequency modulation and amplitude modulation, J.A.S.A., 34, 8, 1962, pp. 1425-1430.
Temporal effects in psychoacoustical excitation, Basic mechanisms in hearing, ed. Möller A.R., Academic Press, New York, 1973, pp. 809-824.

ZWIRNER, E., Phonologische und Phonometrische Probleme der Quantität, Proc. 3rd Int. Cong. Phon. Sci. Ghent., 1938, pp. 58-65.

WOODROW, H., The reproduction of temporal intervals, J. Exp. Psychol., 13, 1930, pp. 473-499.

TABLEAU I

	1° séance voyelle	2° séance son-pur	3° séance son-pur	4° séance voyelle	Moyenne son-pur	Moyenne voyelle	Moyenne générale
Sujet 1	8,1	3,9	5,9	5,7	4,9	6,9	5,9
Sujet 2	5,9	9,9	7,7	5,7	8,8	5,8	7,3
Sujet 3	9,9	7,7	9,9	4,3	8,8	7,1	8,0
Sujet 4	9,7	11,5	5,9	6,5	8,7	8,1	8,4
m	8,4	8,3	7,4	5,6	7,8	7,0	7,4

EXPERIENCE I

Valeurs individuelles minimales de l'indice de sensibilité différentielle $dh \times \frac{100}{0,2}$ pour les 4 séances expérimentales

TABLEAU II

	Son-pur	%	Voyelle	%	% moyen	X^2	Seuil
Sujet 1	189	73,8	200	78,1	76,0	1,29	0,2
Sujet 2	197	77,0	187	73,0	75,0	1,04	0,3
Sujet 3	190	74,2	198	77,3	75,8	0,96	0,3
Sujet 4	199	77,7	197	77,0	77,3	0,04	0,8
Sujet 5	195	76,2	191	74,6	75,4	0,17	0,6
m		75,8		76,0			

EXPERIENCE II

Nombre de réponses correctes (sur 256)
suivant la nature du stimulus.
Résultats individuels.

TABLEAU III

	Son-pur	%	Voyelle	%	% moyen	X^2	Seuil
Sujet 1	99	77,3	101	78,9	78,1	0,09	>0,7
Sujet 2	102	79,7	100	78,1	78,9	0,09	>0,7
Sujet 3	103	80,5	102	79,7	80,1	0,02	>0,8
Sujet 4	103	80,5	105	82,0	81,3	0,10	>0,7
Sujet 5	101	78,9	93	72,7	75,8	1,36	>0,2
m		79,4		78,3	78,8		

EXPERIENCE II

Nombre de réponses correctes (sur 128)
pour les stimuli précédés par un stimulus homogène.
Résultats individuels.

TABLEAU IV

	Son-pur	%	Voyelle	%	% moyen	X^2	Seuil
Sujet 1	73	28,5	168	65,6	47,1	70,75	$<10^{-4}$
Sujet 2	83	32,4	195	76,2	54,3	98,73	$<10^{-4}$
Sujet 3	87	34,0	149	58,2	46,1	31,12	$<10^{-4}$
Sujet 4	125	48,8	187	73,0	60,9	31,54	$<10^{-4}$
Sujet 5	123	48,0	145	56,6	52,3	3,79	$\simeq 0,05$
m		38,4		65,9	52,1		

EXPERIENCE II

Nombre de réponses "long" (sur 256)
suivant la nature du stimulus.
Résultats individuels.

9èmes JOURNEES D'ETUDE SUR LA PAROLE

LANNION 31 mai - 2 juin 1978

SEUILS DIFFERENTIELS ET PERCEPTION DE MODIFICATION D'INTENSITE DANS LA PAROLE CONTINUE

C. SORIN CNET - LANNION

.**RESUME**

Il s'agit dans l'expérience décrite ci-joint, de déterminer les seuils de discrimination d'un accroissement linéaire de pression en fin de phrase. La durée sur laquelle porte la modification varie de 200 à 600 ms par pas de 200 ms.

Pour chacune de ces durées, l'amplitude finale en décibel de la modification nécessaire pour être juste perçue (seuil différentiel relatif d'intensité) est mesurée sur un corpus de 12 phrases. 7 sujets ont participé à l'expérience. L'évolution de ces seuils en fonction de la durée de la modification apportée suit une courbe décroissante atteignant pour des durées comprises entre 400 et 600 ms une valeur stable, indépendante de la durée.

Le type de mécanisme mis en jeu par le système auditif dans cette expérience est discuté. On propose l'hypothèse d'une intégration par l'oreille de la partie finale du signal à partir de la dernière voyelle accentuée précédant le début de la modification.

DIFFERENCE LIMEN AND INTENSITY PERCEPTION OF FLUENT SPEECH

. SUMMARY

The aim of this study is to obtain some partial answers to the following two questions :

- What is the value of difference limen for intensity at the end of sentences, in fluent speech ?
- What temporal integration takes place in this task ?

In this experiment the speech signal (sentences) is modulated by a signal described in figure 1. The amplitude of the modulation at the offset of the sentence is G. The values of the duration T on the increasing portion is 200, 400 and 600 ms. For each value of T, that value of G is measured at the threshold (just notable difference - JND) for 12 sentences presented in pair as "Natural ws Modified". Seven subjects participate at the experiment. As comparaison, the same experiment is reproduced using white noise instead of speech.

The variation of the limen as fuction of the duration T is shown in figures 5 and 6. G decreases when the duration T increases and reaches to a stable value when the duration T becomes greater than 400 ms.

The JND observed cannot be explained with the physical measurements of the difference in intensity at the maixima of the final part of the 2 pair sentences (Fig. 8) not with those of the difference in the mean intensity of the total duration of the modification (Fig. 7).

On the other hand, the JND can be predicted from an integration of the final par of the signal including the last accentuated vowel preceding the beginning of the modification (Fig. 9, 10).

In conclusion, the JND in terms of the value of G are :

- 5 dB ± 1,7 dB for T = 200 ms
- 3 dB ± 1 dB for T = 400 ms
- 2 dB ± 0,8 dB for T = 600 ms,

The corresponding intensity difference limen ΔN are :

- 2,5 dB ± 0,8 dB for T = 200 ms
- 1,3 dB ± 0,3 dB for T = 400 ms
- 1 dB ± 0,2 dB for T = 600 ms.

For the duration T = 400 and 600 ms this values are similar to the "classical" intensity difference limen for stationnary tones or white noise (1 - 1,5 dB). Such value of ΔN is also found for the duration T = 200 ms after the integration as described above.

The comparaison of these results for the sentences with those observed on white noise stimuli shows that the mechanisms of intensity perception in speech cannot be deducted directly from the mechanisms observed for stationnary stimuli such as pure tone, white noise or synthetic vowel.

INTRODUCTION

Les études concernant la perception de sonie des sons stationnaires sont nombreuses et ont abouti à un certain nombre de schémas fonctionnels et à l'élaboration de méthodes de mesure objective (ZWICKER 1960). Par contre, les mécanismes mis en jeu dans la perception de la sonie d'un signal complexe comme le signal de parole restent encore peu étudiés. Les recherches ont porté essentiellement sur la mesure de la sonie globale de phrases (FASTL (1977), Proc. FASE (1975), SORIN (1976)) ou de voyelles isolées (ROSSI 1971).

L'expérience décrite ci-dessous concerne l'étude de la discrimination d'une modification linéaire croissante de pression appliquée sur les 200, 400 et 600 dernières millisecondes de phrases naturelles.

Ce type de modulation a souvent été étudié sur des sons purs. A.M. SMALL (1977) met en évidence qu'une pente de 0,4 dB/sec appliquée à la totalité d'un signal de 2 s de son pur (500 Hz) est juste perçue. D. WOLSK (1964) relève que, sur un son pur, "un accroissement de pression de 150 dB/min (2,5 dB/s) est perçu au bout de 300 ms". Il suggère l'existence possible de deux mécanismes distincts lors de la discrimination de sonie : l'un correspondant à une perception continue de variation d'intensité, l'autre revenant à une comparaison entre la sonie finale et la sonie initiale, faisant donc intervenir un processus de mémorisation. ROSSI (1976) a étudié l'influence d'une modulation de même type sur la sonie globale d'une voyelle synthétique. Il conclut que la sonie résultante correspond à l'intensité du signal aux 2/3 de la durée de la modification.

L'étude de la discrimination d'une telle modulation appliquée au signal de parole avait pour but de donner des éléments de réponse à 2 questions :

* quel est le seuil différentiel d'intensité discriminable en fin de phrases sur de la parole continue ?
* quel mécanisme d'intégration temporelle est mis en jeu dans cette tâche ?

I - PLAN D'EXPERIENCE

I-1 : Mise en forme des signaux

La modification de pression étudiée est représentée sur la fig 1

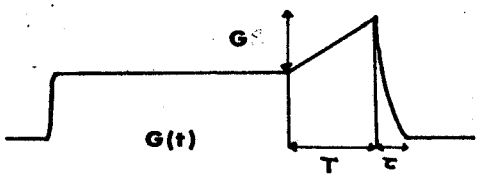


Fig 1 : Type de modulation

Modulation signal

Un appareil spécifique réalisé au laboratoire (R. DAGORNE) permet de multiplier le signal de parole par le signal continu (A). Ce signal peut être ajusté à 20 ms près pour avoir une durée identique à celle du signal de parole, durée à laquelle s'ajoute au temps de décroissance de 40 ms. L'amplitude finale G (gain) de la modification est variable par pas de 1 dB, la durée T de la modification est réglable par pas de 200 ms.

A titre d'exemple la figure 2 montre les oscillogrammes d'une phrase naturalle et de la même phrase "modifiée" et la figure 3, l'évolution comparée de la puissance à court terme (intégrée sur 60 ms) entre une phrase naturelle et la même phrase "modifiée".

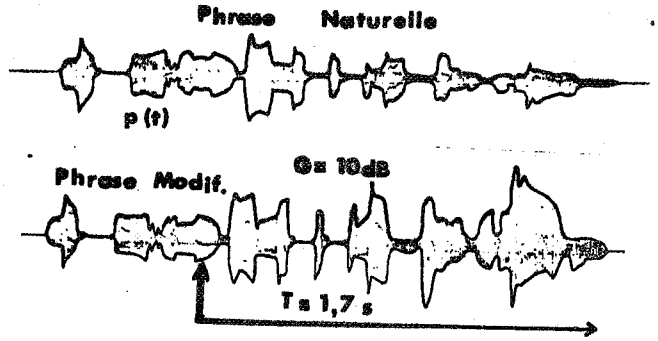


Fig 2 : Oscillogrammes de la phrase "L'été renouvelle la population du village"naturelle et "modifiée"

Sentence "natural" and "modified"



Fig 3 : Evolution comparée de la puissance moyennée sur 60 ms pour la phrase "Nous sommes n'import"où" naturelle et "modifiée" (G = 2 dB, T = 600 ms)

Compared power of one sentence natural and modified

I-2- : Dispositif experimental :

Le dispositif expérimental est représenté sur le schéma ci-dessous :

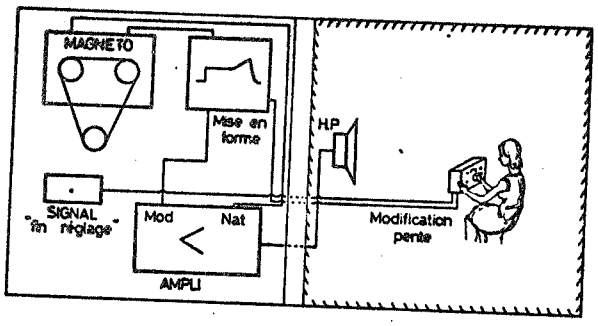


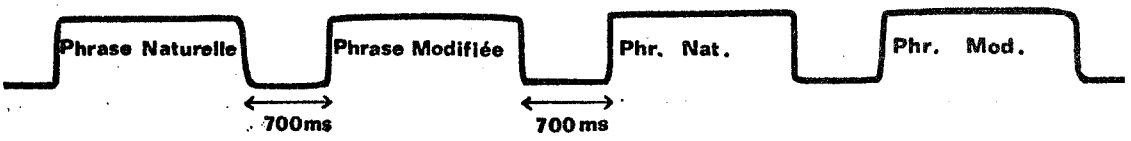
Fig 4 : Dispositif expérimental

Block diagram of apparatus

Le sujet, installé dans une cabine audiométrique, reçoit les signaux par un haut parleur frontal situé à 1 m de lui. Le niveau d'écoute est d'environ 70 dB. Un potentiomètre non gradué permet au sujet de faire varier l'amplitude G de la modification appliquée, de 0 à 10 dB par pas de 1 dB, avec une position G = 20 dB, de "contrôle". A l'aide d'un commutateur commandant un voyant lumineux, le sujet avertit l'expérimentateur de la fin de son réglage. L'expérimenteur est maître de la durée de la modification et du choix des phrases présentée au sujet.

I-3 : Méthode expérimentale

La mesure de discrimination a été effectuée par la méthode des limites. Les phrases étaient présentées par paires et en boucle suivant le schéma ci-dessous.



Ce type de présentation évite qu'un effet d'ordre apparaisse. Chaque mesure comportait 2 phases :

phase 1 : détermination du seuil "descendant"
phase 2 : détermination du seuil "ascendant"

Dans la 1ère phase, le sujet, partant de la modification G = 20 dB, avait pour consigne d'atteindre à l'aide du potentiomètre et par paliers decendants la valeur de G pour laquelle il ne percevait plus de différence entre la phase modifiée et la phase naturelle.

Dans la 2ème phase, le sujet avait pour point de départ la position 0 dB du potentiomètre et comme consigne de régler par paliers ascendants le niveau G de la modification nécessaire pour qu'il puisse indiquer en toute certitude laquelle des deux phases était modifiée.

Le temps nécessaire au sujet pour donner sa réponse n'était pas limité.

Pour chaque sujet, une préexpérience portant sur 2 phases a été réalisée afin de le familiariser avec la procédure proposée. Notons qu'il n'a pas été demandé aux sujets d'évaluer ou de comparer la sonie globale des phrases. La question ne portait que sur l'existence ou non d'une modification perceptible en fin de phrase .

I-4 - Corpus et choix des sujets

6 phrases, enregistrées chacune par 2 locuteurs (1 homme et 1 femme) soit 12 phrases ont été étudiées : (M) (F)

M_1 - F_1 : Les habitants manifesteront dans les rues de la ville
M_2 - F_2 : Vous pensez recommencer l'expérience ?
M_3 - F_3 : C'est ce que vous vouliez savoir ?
M_4 - F_4 : L'été renouvelle la population du village
M_5 - F_5 : Le facteur fait sa tournée à bicyclette
M_6 - F_6 : Nous sommes n'importe où.

Les durées de ces phrases varient de 1,1 à 2,4 s.

7 sujets (3 F et 4 M), audiologiquement normaux et non spécialement entraînés ont participé à l'expérience.

I-5 - Déroulement de l'expérience :

3 durées de modifications ont été étudiées : 200, 400, 600 dernières ms. En effet, une préexpérience avait permis de constater que le seuil de discrimination restait stable et indépendant de la durée de la modification quand celle-ci dépassait 600 ms. Toutefois, à titre de vérification, l'expérience a été poursuivie pour des durées de modification de 1000 ms, sur 3 sujets.

Chaque mesure, comportant l'évaluation du seuil par la méthode des limites a été répétée 3 fois pour chaque sujet, soit un plan d'expérience du type : $D_{(3)} \times P_{(6)} \times L_{(2)} \times R_{(3)} \times S_{(7)}$

D = durée, P = phrase, L = locuteur , R = répétition, S = sujet.

Les couples de phrases "Naturelle/Modifiée" étaient présentés en ordre aléatoire ainsi que les durées de modification. Chaque séance ne dépassait pas 15 mn.

En outre la même expérience a été reproduite sur un signal de 2 s de bruit pseudo-blanc, pour les 3 durées de modification, la mesure n'ayant pu être effectuée que 2 fois par sujet.

II - RESULTATS :

Dans cette partie on appelle "seuil" la valeur liminaire G de la modification.

II-1 - Résultats sur les signaux de bruit blanc :

Le seuil moyen de l'ordre de 1,3 dB (σ = 0,6 dB) apparaît comme étant à peu près indépendant de la durée de la modification. Cette valeur correspond à une augmentation moyenne de niveau du même ordre de grandeur que le seuil différentiel d'intensité sur un signal de bruit blanc observé lors de diverses expériences (ZWICKER 1967 - p. 97). Elle est identique également au seuil différentiel d'intensité observé par WOLSK (1964) sur des sons purs modulés sinusoïdalement en amplitude.

Ces résultats obtenus sur des signaux de bruit blanc nous permettront d'examiner si les mécanismes mis en jeu sont identiques lors de la perception de signaux stationnaires ou de signaux de parole.

II-2- Résultats sur les phrases

Nous avons porté sur la figure 5 les seuils obtenus par chaque sujet, moyennés sur les 12 phrases, pour les 3 durées de modification.

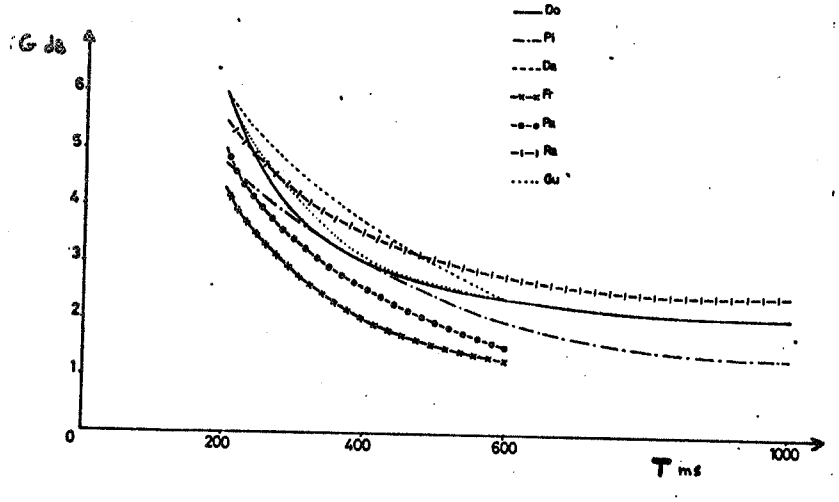


Fig 5 : Valeur G de la modification liminaire en fonction de la durée T pour les 7 sujets (moyenne sur 12 phrases)

The value of the difference limen G as function of the duration for 7 subjects (mean on 12 sentences)

L'évolution de ces seuils en fonction de la durée suit une courbe décroissante atteignant vers 600 ms la valeur de seuil différentiel d'intensité observé habituellement sur un signal de parole (1-2 dB).

Les résultats fournis par 3 sujets pour une durée de 100 ms confirment que les seuils évoluent peu au delà de 600 ms.

Sur le tableau 6 on trouve pour chaque phrase et chaque durée la valeur du seuil moyenné sur tous les sujets.

N° Phrase	200ms	400ms	600ms
M_1	4 dB	2,2 dB	1,3
M_2	4,8	3,2	2
M_3	4,7	2,5	2
M_4	6,2	3	1,7
M_5	3,2	1,8	1,8
M_6	5,7	3,6	3
F_1	7	3,4	2,6
F_2	5,2	4	2,5
F_3	4,6	3	2,2
F_4	5,7	3	2
F_5	6	3,4	2,5
F_6	5	3	2
σ interindividuel sur chaque phrase	1,4 dB	1,2 dB	1 dB

Tableau 6 : Valeur G de la modification liminaire en fonction de T pour les différentes p phrases.

The valeurof the difference limen G as function of the duration T for 12 sentences

Une analyse de variance (programme ANVA 5) portant sur les 3 facteurs durée, phrase, locuteur donne les résultats suivants :

Source	df	F	P
Locuteur	1	15,9	1 %
Phrase	5	1,7	25 %
Durée	2	156	1 %

Soit un effet locuteur significatif ainsi que l'effet durée. Un test "t" confirme que les seuils moyennés sur toutes les phases et tous les locuteurs sont significativement différents pour les 3 durées T :

T	200 ms	400 ms	600 ms
Moy	5,2 dB	3 dB	2,1 dB
σ	1,7 dB	1,1 dB	0,8 dB

Ecart 200 - 400 ms : t=3,8 significatif à 99,8 % (dl = 22)
400 - 600 ms : t=2,5 " à 95 % (dl = 22)

Une analyse plus précise des résultats montre que les phrases donnant lieu à un effet locuteur sont M_1 F_1 M_5 F_5

En effet, une analyse de variance portant sur les 8 phrases restantes donne les résultats suivants :

Source	df	F	P
Locuteur	1	0,03	25 %
Phrase	3	1,31	25 %
Durée	2	100	1 %

En conclusion, cette simple observation des résultats montre que contrairement à ce qui a été observé sur les signaux de bruit, les seuils varient en fonction de la durée T. L'oreille ne semble donc pas être sensible uniquement à la valeur finale de la modification ou à sa pente. Le facteur durée intervient. Il y a donc lieu d'étudier quel mécanisme d'intégration ou de mémorisation est déclenché.

III - MESURES PHYSIQUES ET TENTATIVES D'EXPLICATION

Le signal de parole n'est pas un signal stationnaire. Ceci a pour conséquence qu'un accroissement de pression atteignant 6 dB par exemple, appliqué aux 200 derniers ms d'une phrase n'aura pas le même résultat au terme d'augmentation d'énergie suivant que la fin de la phrase présente par exemple un ou plusieurs pics d'intensité.

C'est pourquoi il est nécessaire de mesurer à quel écart de puissance réel sur le signal de parole correspondent les valeurs de seuils relevées ci-dessous.

Ceci doit nous permettre d'étudier s'il existe une mesure physique simple capable de rendre compte des résultats subjectifs observés. Etant donné que la tâche demandée aux sujets était la même (trouver un seuil) quelle que soit la phrase et la durée de la modification, nous attendons de cette mesure qu'elle donne des résultats semblables Xo quelque soit la phrase et la durée de la modification nous permettant ainsi d'affirmer : "l'oreille effectue ce type de mesure ; si le résultat de cette mesure dépasse la valeur Xo, une différence est perçue entre la phrase modifiée et la phrase non modifiée ; sinon, aucune différence n'est perçue".

Pour chaque phrase et chaque durée T, nous avons relevé le seuil subjectif obtenu (tableau 6) puis calculé l'écart ΔN en niveau de puissance (ou d'énergie) correspondant à cette modification liminaire :

$$\text{soit } \Delta N(T) = 10 \log \frac{\int_0^T G^2(t)p^2(t)dt}{\int_0^T p^2(t)dt} \qquad T = 200, 400 \text{ ou } 600 \text{ ms} \qquad (1)$$

La mesure de ΔN revient à calculer l'augmentation constante d'intensité qu'il faudrait appliquer sur la durée T pour avoir une augmentation d'énergie identique à celle créée par la modification étudiée.

Ces mesures ont été effectuées à l'aide d'un codeur relié à un calculateur T 1600.

Les résultats obtenus sont relevés dans le tableau 7.

Les valeurs ΔN restent voisines de 1 dB pour T = 400 ms ou 600 ms. On retrouve la valeur "classique" d'un seuil différentiel d'intensité sur un signal de bruit, pour des sujets non entraînés. Par contre les valeurs de N pour T = 200 ms varient notablement en fonction de la phrase et sont significativement différents des valeurs ΔN pour T = 400 ou 600 ms (test "t" significatif à 99,8%).

Ceci nous permet donc d'exclure l'hypothèse d'une simple intégration du signal par le système auditif "calquée" sur la durée totale de la modification : si le mécanisme mis en place correspondait à une simple mesure de N, on devrait trouver des valeurs de ΔN de l'ordre de 1 dB, quelle que soit la durée de la modification liminaire, l'oreille détectant une différence entre 2 signaux dès que ΔN dépasse cette valeur.

N° Phrase	200ms	400ms	600ms
M_1	2,1 dB	1,2 dB	0,7
M_2	2,3	1	0,7
M_3	2,1	1,2	1,1
M_4	1,7	1,3	1,1
M_5	2,7	1,8	1,3
M_6	4,5	1,2	1,2
F_1	1,9	1,7	1,2
F_2	2,1	1,0	1,2
F_3	2,7	1,4	1,1
F_4	1,7	1,1	1,1
F_5	2	1,6	1,4
F_6	3,2	0,8	0,8
Moy	2,4 dB	1,3 dB	1 dB
σ	0,8 dB	0,3 dB	0,2 dB

Tableau 7 : Valeur de ΔN pour chaque phrase, pour la valeur liminaire de A.

Difference of mean power ΔN between the modified and the natural sentence in the limen (the definition of ΔN is in equation (1))

On peut alors envisager deux autres hypothèses :

1e hypothèse : l'oreille juge sur les écarts en puissance bien localisés :
(1a) - sur des maxima comme le propose FASTL (1977)
(1b) - au 2/3 de la pente comme le propose ROSSI (1976)

2e hypothèse : l'oreille intègre le signal sur des durées variables, fonction du contenu acoustique ou phonétique du signal et effectue des mesures de type ΔN sur ces durées.

Afin de tester l'hypothèse (1a) nous avons pour chaque phrase et chaque durée relevé la puissance maximale sur la durée T. L'écart ΔN max en dB a ensuite été calculé entre le point de puissance maximale ainsi localisé pour la phrase naturelle et le point correspondant de la phrase modifiée pour la valeur liminaire de la modification.

Les valeurs obtenues sont relevées dans le tableau 8.

N° Phrase	200ms	400ms	600ms
M_1	2,7 dB	0,9 dB	0,5 dB
M_2	2,3	1	0,3
M_3	1,2	1,2	1,2
M_4	0,4	1,3	1
M_5	2,8	2	1,1
M_6	3,9	1	1,6
F_1	0,55	2,5	0,9
F_2	3,1	0,7	1,3
F_3	1	1,2	0,7
F_4	0,3	1	1,1
F_5	0,05	1,5	1,7
F_6	3,2	0,6	0,8

Tableau 8 : Ecart ΔN max en puissance maximale, pour la valeur liminaire de la modification.

Difference of maximal power ΔN max in the limen.

On remarque que les écarts ΔN max varient encore plus que les écarts ΔN relevés précédemment. Si l'oreille effectuait une simple comparaison en puissance des maxima nous aurions dû trouver des écarts ΔNmax semblables pour toutes les phrases et toutes les durées étant donné que pour chaque cas il s'agissait de la mesure du seuil de discrimination. Cette hypothèse ne peut donc être retenue comme seule explication du phénomène.

En ce qui concerne l'hypothèse 1b, celle-ci pose un problème de mesure lié à la forte variabilité du signal de parole. Si un mécanisme identique à celui observé par ROSSI (1976) intervient dans cette tâche de discrimination, les rapports des puissances relevés au 2/3 des durées T entre la phrase modifiée d'une valeur liminaire et la phrase naturelle devraient être constants et indépendants de la phrase et de la durée. Or, ces rapports, que nous avons calculés, varient autant pour les ΔN... Ceci s'explique par le fait que la forte non stationnarité du signal de parole perturbe profondément l'allure de la modification a priori linéaire. En particulier, la localisation auditive exacte du début de la modification dépend fortement de l'intensité du signal à cet instant et ceci suffit à rendre peu probable l'hypothèse d'un traitement identique à celui observé sur des signaux stationnaires.

La 2ème hypothèse a été étudiée de la façon suivante :
Nous avons cherché sur quelle durée il était nécessaire de calculer la puissance moyenne pour que l'écart ΔN moy entre la phrase modifiée sur 200 ms et la phrase naturelle soit comparable au ΔN "classique", de 1 - 1,5 dB, relevé sur des sons stationnaires et observé dans notre cas pour les modifications de durée 400 et 600 ms.

On trouve dans le tableau 9 les écarts ΔN moy mesurés, dans le cas de la modification liminaire de durée 200 ms, sur des durées T_i (temps d'intégration) allant de 200 à 400 ms.

Modification T = 200 ms

Sujets / Durée d'intégration	200 ms	220 ms	250 ms	300 ms	400 ms
M_1	2,1 dB			*	1,1
M_2	2,3		1,5 *	0,2	0,1
M_3	2,1			*	1,3
M_4	1,7	* 1	0,7	0,4	0,2
M_5	2,7		2,4	*	1,3
M_6	4,5			3,5 *	0,8
F_1	1,9			*	1
F_2	2,1	2,0 *	0,4	0,1	0,1
F_3	2,7	1,5 *	1,1	0,7	0,3
F_4	1,7	1,6 *	0,5	0,3	0,1
F_5	2	*	0,9	0,2	
F_6	3,2		2,2 *	0,4	

Tableau 9 : Valeurs de ΔN moy pour la valeur liminaire G de la modification T = 200 ms et différentes durées d'intégration T_i

$$\Delta N \text{ moy } (T_i) = 10 \log \frac{\int_0^{T_i} G(t) p^2(t) dt}{\int_0^{T_i} p^2(t) dt} \quad (2)$$

Difference of mean power ΔN moy ((2)) in the limen for different value of integration time T_i (duration of modification T = 200 ms).

* = Valeur de T_i pour laquelle N moy est compris entre 1 et 1,5 dB

On a marqué d'un astérisque la localisation de la durée d'intégration critique ainsi définie.

Nous avons ensuite relevé la localisation temporelle des maxima d'intensité en fin de phrase. Sur la figure 10 sont reportées les durées critiques tirées du tableau 9 (*) et les localisations temporelles des maxima (o)

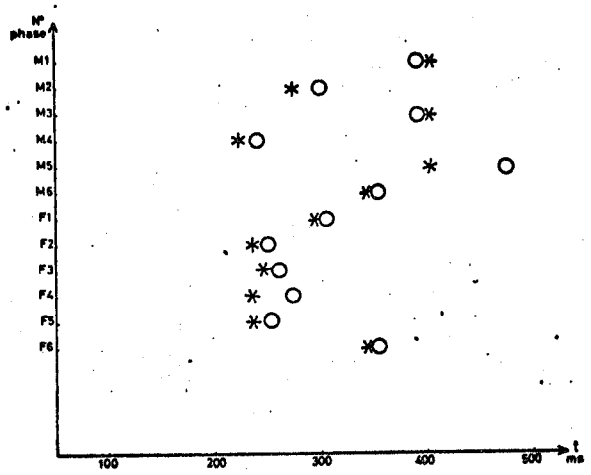


Fig 10 : Localisation des maxima d'intensité (o) et des points d'égalisation des N_{moy} (*)

Localisation of the intensity maxima (o) and the critical durations T_i (*) derived from Table 9

L'écart entre les maxima et les points d'égalisation des ΔN_{moy} est très réduit. La 2° hypothèse doit donc être retenue : dans cette tâche de discrimination les mécanismes mis en jeu correspondent à une comparaison de puissance moyenne sur une durée au moins égale à 200 ms intégrant la partie finale du signal en incluant le dernier maxima.

L'analyse phonémique nous permet d'identifier comme suit les différents maximas :

Phrase 1 : de la ville ↑

Phrase 2 : expérience ↑

Phrase 3 : savoir ↑

Phrase 4 village ↑

Phrase 5 a bicyclette ↑

Phrase 6 n'importe où ↑

La deuxième hypothèse revient donc, dans le cadre de ce corpus, à supposer, que lors de la discrimination d'une modification d'intensité en fin de phrase, tout se passe comme si l'oreille intégrait le signal à partir de la voyelle la plus intense précédant le début de la modification (ou située à plus de 200 ms de la fin du signal).

IV - CONCLUSION

Cette expérience nous permet d'affirmer qu'un accroissement linéaire de pression est perçu une fin de phrase dès que l'amplitude finale de cette variation atteint :

* 5dB ± 1,7dB lorsque la durée de la modification est de 200 ms
* 3dB ± 1dB " " 400 ms
* 2dB ± 0,8dB " " 600 ms

Ces valeurs correspondent à une augmentation d'intensité supposée uniforme sur la durée de la modification de :

* 2,5dB ± 0,8dB sur 200 ms

* 1,3dB ± 0,3dB sur 400 ms

* 1dB ± 0,2dB sur 600 ms.

Le seuil de discrimination n'est pas fonction uniquement de la pente mais varie suivant la durée.

Pour des accroissements portant sur plus de 400 ms, on retrouve des seuils correspondant à une modification uniforme de niveau identique au seuil différentiel d'intensité classique sur des sons stationnaires.

Malgré la taille réduite du corpus, il semble que l'hypothèse d'une intégration par le système auditif du signal à partir de la voyelle la plus intense précédant le début de la modification (ou située à plus de 200 ms de la fin) puisse être retenue.

Ceci suggère ainsi que nous l'avions supposé lors d'expériences antérieures (SORIN 1976) que les mécanismes de traitement de l'intensité sur un signal de parole ne peuvent être déduits directement des mécanismes observés lors de la perception de sons stationnaires.

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

FASTL H. (1977) — Loudness of running speech
J. Audiol. Technique 16, 2-13

Proc. FASE (1975) — Thème 2 : La sonie de la parole et de la musique
Paris

ROSSI M. (1971) — L'intensité spécifique des voyelles
Phonetica 24, 3, 129-161.

ROSSI M. (1976) — La perception des modulations d'intensité sur les voyelles. Travaux de l'Institut de Phonétique d'AIX III - 362-457

SMALL A.M. (1977) — Londness perception of signals of monotonically changing sound pressure
JASA 61, 5, 1293-1298

SORIN C. (1976) — Etude de la sonie de la parole : quelques expériences préliminaires
Recherches Acoustiques III 115-126 CNET-LANNION

WOLSK D. (1964) — Discrimination limen for loudness under varying rates of intensity change
JASA 36, 7, 1277-1282

ZWICKER E. (1967) — Das Ohr als Nachrichtenempfänger
Hirzel Verlag - STUTTGART

9èmes JOURNEES D'ETUDE SUR LA PAROLE

LANNION 31 mai - 2 juin 1978

LES INDICES ACOUSTIQUES ET PERCEPTUELS DES CONSONNES CONSTRICTIVES DU FRANCAIS APPLICATION A LA SYNTHESE

Michel CHAFCOULOFF et Albert DI CRISTO

Institut de Phonétique d'Aix

.**RESUME**

Nous nous proposons, dans ce travail qui porte sur les constrictives du français, de vérifier par des expériences de synthèse la valeur des indices acoustiques (caractéristiques spectrales et microprosodiques) que nous avons dégagés dans des recherches antérieures (CHAFCOULOFF, DI CRISTO et SEIMANDI, 1976, DI CRISTO et CHAFCOULOFF, 1978a, 1978b).

Notre étude se fonde principalement sur deux expériences. La première consiste à générer, en réutilisant tous nos résultats, des séquences VCV. Ces dernières sont présentées à des sujets qui doivent identifier la consonne intervocalique. Dans la seconde, qui est destinée à tester l'importance des effets du contexte, nous substituons, par exemple, un [ʃ] arrondi à un [ʃ] non arrondi dans le contexte de ce dernier.

Les taux d'identification très élevés obtenus dans la première expérience confirment la validité de la méthode d'analyse adoptée et montrent que l'intégration à la synthèse des caractéristiques microprosodiques rend cette dernière plus performante. Les résultats de la seconde expérience montrent que la perception catégorielle des consonnes est grandement affectée si on ne tient pas compte des effets du contexte. Ils permettent, en outre, d'évaluer le rôle respectif des pôles de tonalité et de stridence pour l'identification des constrictives du français.

SUMMARY

The aim of this work is to test, through experiments in synthesis, the perceptual relevance of acoustic cues previously outlined in a recent study on fricative consonants (CHAFCOULOFF, DI CRISTO, SEIMANDI, 1976). It was shown that the principal acoustic parameters (duration, intensity, frequency) are influenced by effects of coarticulation. The magnitude of microprosodic variations was also investigated (DI CRISTO et CHAFCOULOFF, 1978a, 1978b). Those data were used in the two experiments of the study.

In the first one, we used stimuli [VCV] produced by a computer-controled formant synthesizer. Listeners were asked to identify the intervocalic consonants. The identification rate was high. The results show :

1) the validity of the multiparametric method.

2) the intelligibility and quality of synthetic speech is improved by the use of microprosodic variations.

The second experiment was aimed to check the importance of effects related to vocalic context (i. e. a synthetic "rounded" [ʃ] was produced in an unrounded vocalic context and vice versa).

The results show that the lack of correspondence between the acoustic structure of the fricative and the adjacent vowel consequently impairs the categorical perception of the consonant.

Moreover , they demonstrate the necessity of a syntagmatic approach in work on synthesis. Finally, they show the perceptual relevance of the tonality pole and the strident pole in the identification of french fricative consonants.

INTRODUCTION

La génération d'une parole artificielle intelligible et de bonne qualité, dépend en grande partie du soin apporté à l'analyse acoustique qui la précède.

Au cours de ces dernières années, de nombreuses études ont été entreprises, qui avaient précisément pour but de définir les indices acoustiques des consonnes constrictives dans diverses langues. Nous nous bornerons à citer ici les principales : celles de HUGHES and HALLE (1959) et de STREVENS (1960), sur l'anglais ; celle de ANDERSEN (1976) sur l'allemand et le danois ; celle de MARTONY and al (1962) sur le suédois ; celle de JASSEM (1968) sur le polonais ; celles de HALLE (1959) et de FANT (1960) sur le russe ; celle de NAKATA (1960) sur le japonais ; et celle de ROSSI (1974) sur l'italien (1).

Comme aucune recherche approfondie n'avait été effectuée en français, nous nous sommes efforcés récemment (CHAFCOULOFF, DI CRISTO et SEIMANDI, 1976), de dresser l'inventaire des indices acoustiques relatifs aux constrictives [f - v - s - z - ʃ - ʒ] et d'évaluer l'influence du contexte vocalique sur la réalisation de ces consonnes.

Nous nous proposons, dans la présente étude, de vérifier la validité des indices découverts, en soumettant à des tests perceptuels les stimuli synthétiques générés d'après les résultats de notre analyse acoustique.

ANALYSE DU PROBLEME

Les deux méthodes les plus fréquemment employées pour la synthèse des constrictives sont les suivantes :

a) On simule un segment constrictif (isolé ou en contexte) à partir d'un bruit blanc filtré de 0 à 10000 Hz. Dans ce cas, les attributs acoustiques de la consonne sont spécifiés en termes de pôles (zones de résonance) et de zéros (anti-résonances). Ce type de synthèse est utilisé notamment par MARTONY and al (1962, p. 212), par HEINZ and STEVENS (1961, p. 593) et par KACPROWSKI and al (1973, p. 231). Bien qu'il permette d'obtenir de façon satisfaisante des consonnes comme : [s], [ʃ] et [ç], il s'avère inefficace pour la production de [f] et [θ].

b) La seconde méthode est fondée sur l'emploi de deux formes de synthèse. C'est ainsi que les chercheurs du groupe HASKINS (DELATTRE and al, 1964) ont utilisé le Pattern Playback pour définir les indices acoustiques propres aux différents sons du langage. Malgré l'apport incontestable de ces recherches, la qualité des consonnes constrictives demeure relativement médiocre, à cause des limites inhérentes à l'appareil employé.

D'autre part, HEINZ and STEVENS (op. cit.) et KACPROWSKI and al (op. cit.) se servent de synthétiseurs à résonance pour générer les spectres simplifiés des fricatives de l'anglais et du polonais. Leur étude se limite, toutefois à la synthèse des consonnes non voisées : [f, θ, s, ʃ]et[f, s, ʃ, ç, x]. La qualité perceptuelle des résultats varie en fonction de la nature de la consonne synthétisée.

Dans cette perspective, le travail que nous présentons ici a un triple objectif :

---

(1) Pour plus de détails, voir CHAFCOULOFF, M. (1976).

1. Préciser dans quelles limites la réutilisation fidèle des variations paramétriques analysées dans l'étude acoustique peut conduire à une perception optimale des consonnes non voisées : [f - s - ʃ].

2. Examiner les problèmes posés par la synthèse des consonnes voisées [v - z - ʒ].

3. Vérifier l'hypothèse suivant laquelle il est indispensable de tenir compte, dans les règles de synthèse, des faits de coarticulation.

PROCEDURE EXPERIMENTALE

Nous avons utilisé le système de synthèse paramétrique FSYNT, élaboré par ABBEG, ESPESSER et TESTON (à paraître). Les différentes phases du processus de génération des stimuli sont représentés sur la figure 1.

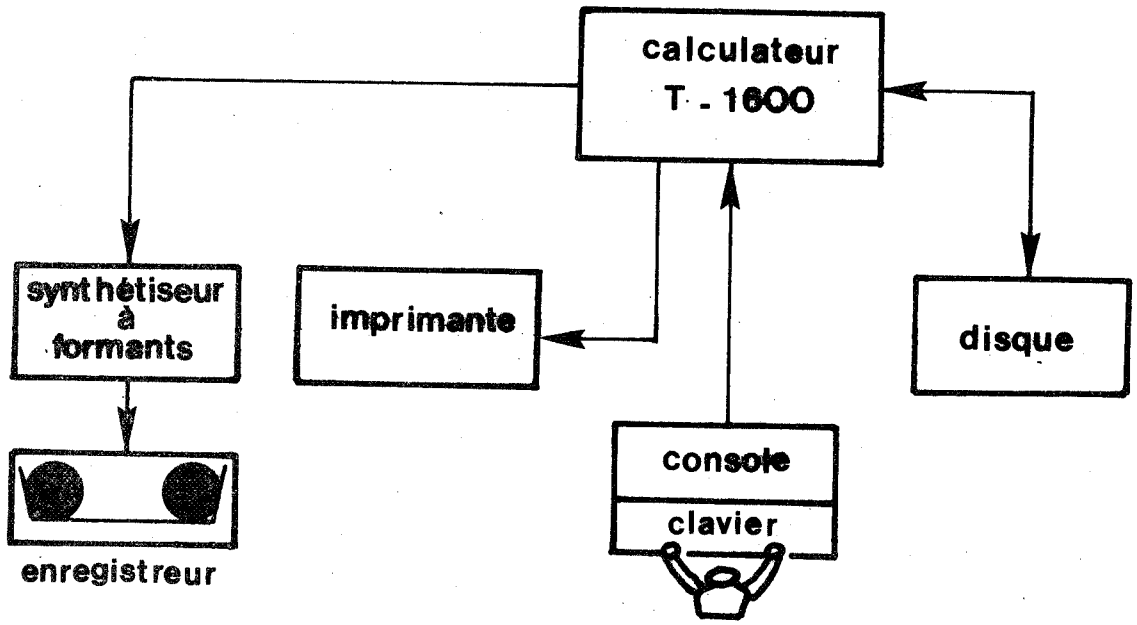


Fig. 1 : Diagramme du plan d'expérience.
General plan of the experiment

L'opérateur entre, à l'aide d'un clavier, les données relatives aux différents paramètres, qui sont alors affichées sur la console de visualisation. Ces données sont traitées, stockées sur un disque, et transmises au synthétiseur, qui les régénère sous forme vocale.

L'opérateur dispose également d'une imprimante, qui lui fournit un état des paramètres sur des listings.

Des treize paramètres programmables du système FSYNT, nous n'avons retenu que les huit suivants :

1.- Fo (fréquence fondamentale).
2-4.- F1, F2, F3 (fréquence des trois premiers formants).
5.- Ao (amplitude du voisement).

6.- Ab (amplitude du bruit).
7-8.- B1, B2 (les deux pôles de bruit).

La largeur de bande des formants est fixée en synthèse manuelle : F1 : 80 Hz, F2, F3 : 100 Hz. Celle des pôles de bruit varie en fonction des fréquences retenues (HEINZ and STEVENS, 1961, p. 594). Pour B1 = 2000-3000 Hz, B1 = 500-800 Hz ; pour B2 = 5000-8000 Hz, B2 = 1000-1500 Hz.

Les valeurs des trois paramètres de base (durée, intensité, fréquence) sont directement issues de notre analyse (op. cit., pp. 74-99).

DUREE

Le pas utilisé est de 5 msec. La durée de V1 varie en fonction du mode articulatoire de C . La durée de la tenue de C est définie en fonction du lieu et du mode d'articulation de la consonne. La durée des transitions est fixée pour [f - s ʃ] à 40 msec et pour [v - z - ʒ] à 50 msec.

INTENSITE

Les valeurs de l'intensité globale de la consonne sont établies en fonction :

a) de la nature de la consonne

C'est ainsi, par exemple, que l'énergie de [f] est inférieure à celle des consonnes [s] et [ʃ].

b) de la nature de la voyelle

Conformément aux résultats de BUSH (1964, p. 51), notre étude montre (op. cit., p. 93) que l'intensité de la consonne est plus grande au contact d'une voyelle diffuse qu'à celui d'une voyelle compacte (voir également : DI CRISTO et CHAFCOULOFF, 1978a).

FREQUENCE

Les caractéristiques intrinsèques des différentes consonnes sont reproduites (DI CRISTO et CHAFCOULOFF, 1978b). En ce qui concerne les variations des pôles de bruit, nous n'avons pas pu, à cause des caractéristiques de l'appareil, synthétiser plusieurs pôles dans les basses fréquences : 0-3000 Hz. En conséquence, un seul pôle bas moyen a été réalisé.

Première expérience

Les stimuli employés sont du type :
V1 [a] + C [f, v, s, z, ʃ, ʒ] + V2 [a]

Un état des différents paramètres retenus est présenté dans le tableau I (voir page suivante).

Les stimuli sont évalués perceptuellement, modifiés si cela est nécessaire et enregistrés sur bande magnétique. Nous avons effectué une analyse spectrographique en filtre large : 80-8000 Hz, ainsi qu'une étude des tracés de Fo et d'intensité (variations microprosodiques), à l'aide d'un détecteur de mélodie et d'un intensimètre couplés à un calculateur T. 1600 (programme MELINT élaboré par ESPESSER, 1978).

ETAT DES PARAMETRES. FICHIER: afa

TP	XX	F0	F1	F2	F3	A0	AN	AB	B0	B1	B2	B2	BV	B1
0	1	50	749	1353	2496	10	??	10	1000	1000	50	2000	0	40
40	????	159	749	1353	2496	30	??	10	????	????	???	????	??	???
90	????	159	749	1353	2496	35	??	10	????	????	???	????	??	???
115	2	50	100	696	2002	20	??	10	????	????	???	????	??	???
125	????	50	100	696	2002	10	??	10	????	????	???	????	??	???
135	????	50	100	696	2002	10	??	25	????	1804	???	2000	??	???
145	????	50	100	696	2002	10	??	35	????	1804	???	2000	??	???
265	????	50	100	696	2002	10	??	25	????	1804	???	2000	??	???
275	????	50	100	696	2002	10	??	25	????	1804	???	2000	??	???
285	????	165	100	696	2002	20	??	10	????	????	???	????	??	???
310	3	151	749	1353	2496	25	??	10	????	????	???	????	??	???
350	????	140	749	1353	2496	35	??	10	????	????	???	????	??	???
460	3	110	749	1353	2496	22	0	10	1000	1000	50	8500	0	40

ETAT DES PARAMETRES. FICHIER: asa

TP	XX	F0	F1	F2	F3	A0	AN	AB	B0	B1	B2	B2	BV	B1
0	1	91	749	1353	2496	10	0	0	1000	1000	50	2000	0	40
40	????	159	749	1353	2496	30	??	??	????	????	???	????	??	???
90	????	159	749	1353	2496	35	??	??	????	????	???	????	??	???
130	2	???	????	???	????	35	??	??	????	????	???	????	??	???
140	????	50	100	1804	2798	10	??	10	????	????	???	????	??	???
150	????	???	????	????	????	??	??	35	????	5196	???	8206	??	???
160	????	???	????	????	????	??	??	35	????	5196	???	8206	??	???
340	????	???	????	????	????	??	??	35	????	5196	???	8206	??	???
350	????	50	100	1804	2798	10	??	35	????	5196	???	8206	??	???
360	????	165	????	????	????	30	??	10	????	????	???	????	??	???
400	3	151	749	1353	2496	40	??	??	????	????	???	????	??	???
480	????	140	749	1353	2496	40	??	??	????	????	???	????	??	???
560	3	110	749	1353	2496	35	0	10	1000	5196	50	8206	0	40

ETAT DES PARAMETRES. FICHIER: aʃa

TP	XX	F0	F1	F2	F3	A0	AN	AB	B0	B1	B2	B2	BV	B1
0	1	159	749	1353	2496	10	??	??	????	????	???	????	??	???
40	????	159	749	1353	2496	30	??	??	????	????	???	????	??	???
90	????	159	749	1353	2496	35	??	??	????	????	???	????	??	???
130	2	50	100	2000	2194	35	??	??	????	????	???	????	??	???
140	????	50	????	????	????	10	??	10	????	????	???	????	??	???
150	????	???	????	????	????	??	??	30	????	2608	???	5206	??	???
160	????	???	????	????	????	??	??	35	????	2608	???	5206	??	???
335	????	???	????	????	????	??	??	35	????	2608	???	5206	??	???
345	????	50	100	2000	2194	10	??	30	????	2608	???	5206	??	???
355	????	165	100	2000	2194	30	??	??	????	????	???	????	??	???
395	3	151	749	1353	2496	35	??	??	????	????	???	????	??	???
480	????	140	749	1353	2496	35	??	??	????	????	???	????	??	???
565	????	110	749	1353	2496	30	??	??	????	????	???	????	??	???

ETAT DES PARAMETRES. FICHIER: ava

TP	XX	F0	F1	F2	F3	A0	AN	AB	B0	B1	B2	B2	BV	B1
0	1	159	749	1353	2496	10	0	10	1000	1000	50	2000	0	40
40	????	159	749	1353	2496	30	??	??	????	????	???	????	??	???
130	????	159	749	1353	2496	25	??	??	????	????	???	????	??	???
180	2	151	351	????	????	20	??	20	????	1804	???	8500	??	???
235	????	129	351	????	????	20	??	20	????	1804	???	8500	??	???
280	????	151	351	????	????	20	??	20	????	1804	???	8500	??	???
330	3	159	749	1353	2496	25	??	10	????	????	???	????	??	???
405	????	140	749	1353	2496	30	??	??	????	????	???	????	??	???
480	????	110	749	1353	2496	25	0	0	????	????	???	????	0	40

ETAT DES PARAMETRES. FICHIER: aza

TP	XX	F0	F1	F2	F3	A0	AN	AB	B0	B1	B2	B2	BV	B1
0	1	91	749	1353	2798	10	??	??	????	????	???	????	??	???
40	????	159	749	1353	2496	30	??	??	????	????	???	????	??	???
130	????	159	749	1353	2798	40	??	??	????	????	???	????	??	???
170	2	156	252	1804	????	35	??	10	????	????	???	????	??	???
180	????	151	252	????	????	30	??	30	????	5196	???	8206	??	???
235	????	129	252	????	????	30	??	30	????	5196	???	8206	??	???
290	????	151	252	1804	????	30	??	30	????	5196	???	8206	??	???
330	3	159	749	1353	2798	35	??	10	????	????	???	????	??	???
390	????	140	749	1353	2496	40	??	??	????	????	???	????	??	???
480	????	110	749	1353	2798	35	??	??	????	????	???	????	??	???

ETAT DES PARAMETRES. FICHIER: aʒa

TP	XX	F0	F1	F2	F3	A0	AN	AB	B0	B1	B2	B2	BV	B1
0	1	159	749	1353	2496	10	??	??	????	????	???	????	??	???
40	????	159	749	1353	2496	30	??	??	????	????	???	????	??	???
130	????	159	749	1353	2496	35	??	??	????	????	???	????	??	???
170	2	156	252	2000	2194	30	??	10	????	????	???	????	??	???
180	????	151	252	????	????	35	??	30	????	2608	???	5206	??	???
235	????	129	252	????	????	35	??	30	????	2608	???	5206	??	???
290	????	151	252	2000	2194	35	??	30	????	2608	???	5206	??	???
330	3	159	749	1353	2496	30	??	30	????	2608	???	5206	??	???
390	????	140	749	1353	2496	35	??	10	????	????	???	????	??	???
480	????	110	749	1353	2496	30	??	??	????	????	???	????	??	???

Tableau I : Etat des paramètres acoustiques : séquences
Table of acoustic parameters.

Chacune des séquences VCV est présentée dix fois, suivant un ordre aléatoire. Soixante stimuli ont été ainsi soumis au jugement de deux groupes de sujets. Le premier est constitué d'autiteurs non entraînés. Le second comprend 9 phonéticiens qui avaient déjà participé à des tests de ce type. Les stimuli sont présentés en champ libre, dans des conditions d'écoute normales (le niveau sonore au niveau de l'oreille des sujets est de 70 dB).

Afin de pouvoir comparer nos résultats à ceux de HUGHES and HALLE (1956, p. 309) et de HEINZ and STEVENS (1961, p. 595), nous avons utilisé le même type de test : on demande au sujet d'identifier la consonne qu'il vient d'entendre. Il convient de préciser que les consonnes incluses dans le test ont fait l'objet d'une présentation préalable.

RESULTATS

Les résultats concernant les deux groupes sont présentés conjointement sur la figure 2. Le test de X2 démontre que les différences entre les groupes de sujets ne sont pas significatives (X_2 (5) = 1.435, $<$ .90). D'autre part, afin d'évaluer la représentativité de l'échantillon par rapport à la population parente, nous avons testé l'hypothèse nulle à l'aide de la formule : $0.50 \pm t_{(.01)} \sqrt{\frac{0.50 \times 0.50}{n}}$. Les pourcentages d'identification sont toujours supérieurs à l'intervalle de confiance, ce qui permet de rejeter l'hypothèse nulle.

Enfin, nous avons établi une matrice de confusion, qui permet d'appréhender globalement la distribution des réponses correctes et des réponses erronées (Tableau II).

1. cas des consonnes [f , s , ʃ]

Ces consonnes dont l'identification repose principalement sur la perception de deux indices : localisation des pôles de bruit et énergie relative des pôles, ne posent aucun problème (identification supérieure à 90 %). En ce qui concerne [f] et [ʃ], les taux d'intelligibilité sont supérieurs à ceux de KACPROWSKI and al (1973, p. 239), pour lesquels les pourcentages ne sont que de 72 % et de 75 %, respectivement. Cette différence s'explique en partie par les confusions notées par ces derniers entre, d'une part, [f] et [s] (23 %) et, d'autre part, [ʃ] et [ç] (18 %). Les confusions entre [ʃ] et [ç] sont dues à la proximité acoustique (et perceptuelle) de ces deux consonnes. Pour notre part, nous n'avons pas relevé de confusion notable entre [f] et [s]. LAMBERT (1954) et TARNOCZY (1954) établissent une relation entre ces deux consonnes, qui ont en commun un pôle de bruit aux environs de 8000-9000 Hz. Ce pôle que l'on considère à juste titre comme le corrélat acoustique du trait de stridence (JAKOBSON and al, 1952, p. 24), est cependant plus intense pour [s] que pour [f]. Combiné à un premier rôle (P1) moyen de 1800 Hz, un P2 plus faible à 8000-9000 Hz, permet d'obtenir une excellente synthèse de [f] et d'éviter toute confusion avec la constrictive [s], pour laquelle on doit nécessairement programmer un pôle de stridence plus intense. Cette observation confirme ainsi les remarques de HALLE (1959, p. 140) et de ROSSI (1974, p. 313) sur le caractère graduel de l'indice de stridence.

2. cas des consonnes [v , z , ʒ]

La synthèse des consonnes voisées [v , z , ʒ] est plus délicate à réaliser. Il s'agit, dans ce cas, de combiner les effets du voisement à ceux des bruits. Nous devons, dans ces conditions, ne pas sous-estimer les défauts de la source de voisement du synthétiseur. Les pourcentages d'identification moins

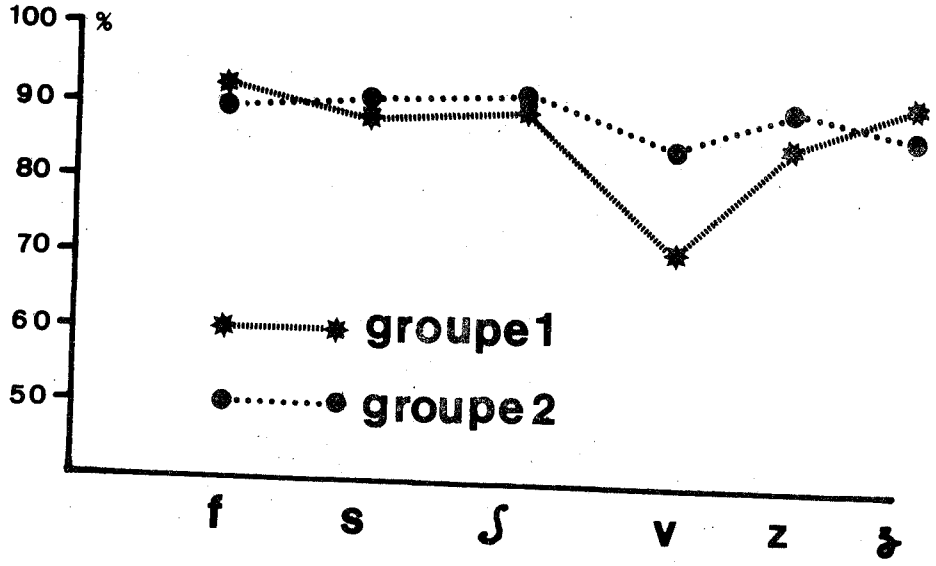


Fig. 2 : Pourcentages d'identification des stimuli synthétiques.

Percentages of identification of synthetic stimuli.

stimuli synthétiques	identification f %	s %	ʃ %	v %	z %	ʒ %
f	92	8	—	—	—	—
s	—	99	—	—	1	—
ʃ	—	—	96	—		4
v	12	—	—	88	—	—
z	—	—	—	—	95	5
ʒ	—	—	—	—	4	96

Tableau 2 : Matrice de confusion (Première Expérience, Groupes 1 et 2).

Confusion matrix (1st experiment : Groups 1 and 2).

élevés reflètent cette difficulté.

Comme [v] est une consonne à faibles bruits (ou dénuée de bruits), son identification repose nécessairement sur d'autres indices (HARRIS, 1958, p. 5). Cet auteur a montré que les consonnes [f , θ , v , ð] sont perçues essentiellement grâce aux transitions de formants. D'après ROSSI (1974, p. 261), on réalise une bonne synthèse de [v] en provoquant une rupture négative de Fo d'environ 20 Hz et une atténuation d'intensité de 14 dB. Malgré la prise en compte de ces indices, la perception de [v] demeure nettement moins bonne que celle des autres consonnes (88 %). En effet, l'identification de cette consonne dépend de plusieurs facteurs. Si l'amplitude du bruit est trop forte, on perçoit un [f]. Si elle est trop faible, on entend soit [β], soit [b]. Si l'amplitude du voisement est trop grande, on crée une structure formantique artificielle, qui conduit à la perception de [w] dont les transitions de formants sont également négatives (DELATTRE, 1968, p. 199). Enfin, si l'amplitude du voisement est trop faible, on élimine les indices contenus dans les transitions de formants et la consonne n'est plus identifiable. Les résultats obtenus pour [v] confirment ceux de PECKELS et ROSSI (1971, p. 9), qui constatent que ce sont précisément les consonnes graves diffuses [f] et [v] qui obtiennent les taux de reconnaissance les plus bas en voix vocodorisée (respectivement : 84 et 7.2 %).

La synthèse des consonnes [z] et [ʒ] dont l'intensité des bruits est plus importante, s'avère plus facile à réaliser. A la suite des travaux de INGEMANN (1960, p. 1501), nous avons constaté, dans notre recherche, que l'introduction de formants de voisement dans les portions de bruit contribue à améliorer l'intelligibilité de ces consonnes. Il convient cependant de préciser que l'amplitude du voisement pendant la tenue de [ʒ] est réduite de 5 dB, par rapport à [z].

Alors que pour [z] les deux formants de faible énergie sont nettement distincts (Figure 3.3.), leur concentration au niveau des locus de F2 et F3 (vers 2000-2200 Hz) pour [ʒ] (Fig. 3.4) donne lieu à la formation d'un seul formant d'amplitude élevée. L'intelligibilité est fortement perturbée dans ce cas, la consonne [ʒ] pouvant même être confondue avec [l], si le niveau de bruit est faible.

Deuxième expérience

Il est bien connu que dans la parole naturelle les phénomènes de coarticulation ont un rôle perceptuel très important (FANT and al (1970), LEHISTE and SHOCKEY (1972), DANILOFF and HAMMARBERG (1973). En ce qui concerne la synthèse, KACPROWSKI and al (1973, p. 241) affirment que pour être perçue de façon optimale, chaque consonne doit avoir deux ou trois variantes, sélectionnées en fonction du contexte vocalique. (cf. également le concept de "modifieur contextuel" présenté par LIBERMAN and al, 1959).

Il nous a semblé intéressant d'évaluer l'importance perceptuelle de ces variations pour les consonnes françaises. Dans ce but, nous avons synthétisé des stimuli du type :

V1 [i] + C [s , ʃ [+ arr]] + V2 [i] (1)

et

V1 [u] + C [s , ʃ [- arr]] + V2 [u]

(1) N.B. [s , ʃ [+ arr]] signifie que la structure acoustique de la consonne est celle d'une constrictive dans le contexte vocalique [+ arrondi] de [u] (voir le tableau III).

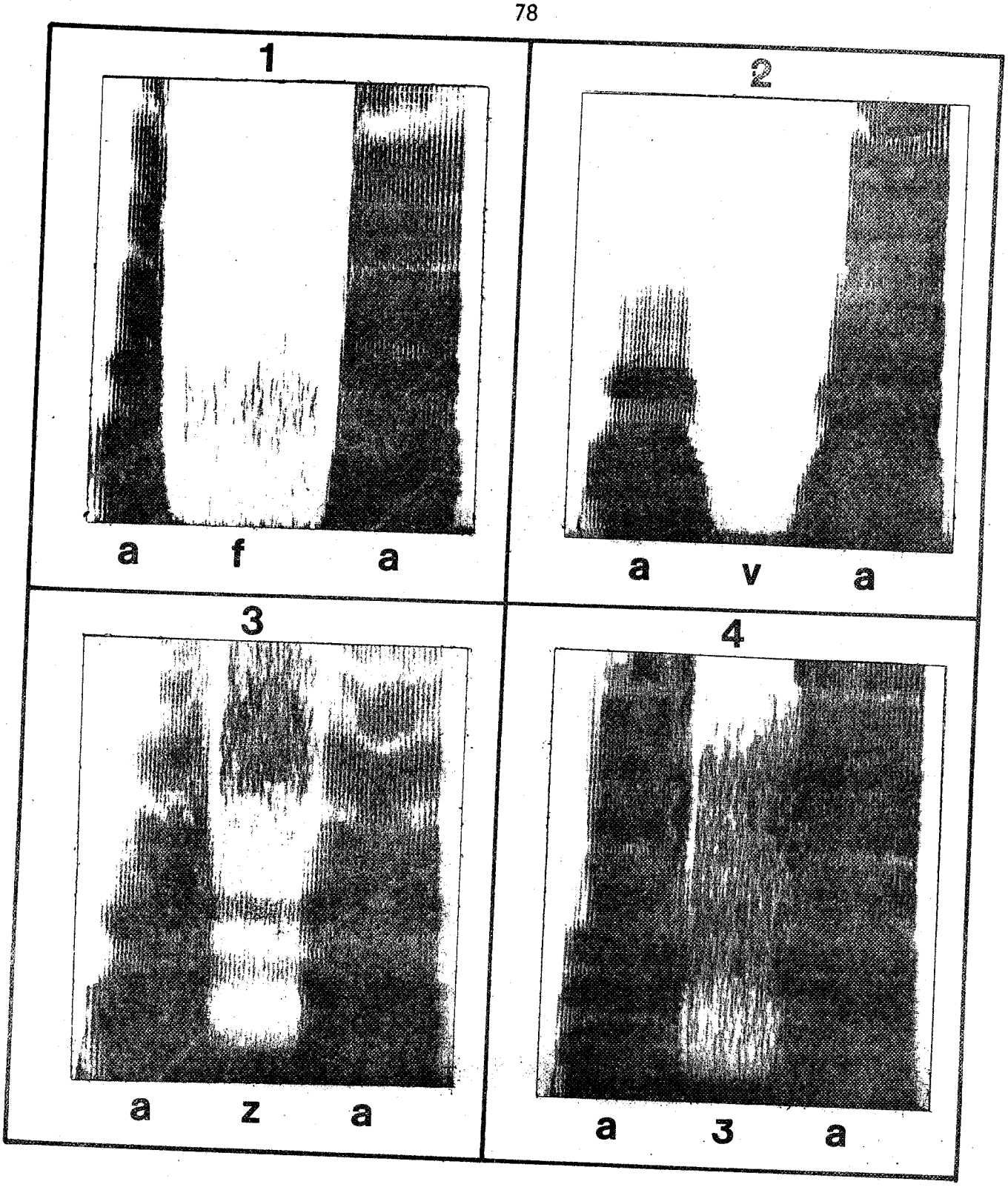


Fig. 3 : Spectrogrammes en filtre large (80-8000 Hz) des séquences synthétiques : [a f a] , [a v a] , [a z a] , [a ʒ a].

Wide-band spectrograms (80-8000 cps) of synthetic sequences [a f a] , [a v a] , [a z a] , [a ʒ a].

S	P1	P2	ʃ	P1	P2
S[i]	5.2	8.2	ʃ[i]	2.6	5.2
S[u]	3.8	8.2	ʃ[u]	1.4	4.6

Tableau III : Valeurs (en Khz) des P1 et P2 de [s] et [ʃ] en contexte [-arr] et [+ arr].

Values (in KHz of P1 and P2 of [s] et [ʃ] between rounded and unrounded vowels.

		identification			
		f %	s %	ʃ %	? %
ʃ	i ʃ[i] i	6	—	94	—
	i ʃ[u] i	84	—	16	—
S	i s[i] i	—	96	2	2
	i s[u] i	4	42	32	4
	u s[i] u	—	98	2	—
	u s[u] u	—	98	2	—

Tableau IV : Matrice de confusion (Deuxième Expérience).

Confusion matrix (2nd experiment).

que nous avons comparés aux séquences normales :

V1 [i] + C [s , ʃ [-arr]] + V2 [i]

et

V1 [u] + C [s , ʃ [+arr]] + V2 [u]

Ces stimuli ont été présentés à 10 sujets dans des conditions identiques à celles du test précédent. La seule différence notable est que les sujets pouvaient répondre par l'incertitude (?).

Les résultats de ce test (tableau IV) montrent à l'évidence que l'identification des consonnes est sérieusement perturbée par ces modifications.

Quand on présente dans un contexte [- arr] le spectre bémolisé de [ʃ], 94 % des sujets perçoivent un [f]. Ce fait s'explique aisément si l'on considère que la plus grande partie de l'énergie est concentrée dans le bas du spectre, le pôle de tonalité se situant alors vers 1400 Hz.

La présentation de [i s^{u} i], entraîne une baisse radicale du pourcentage d'identification de [s], qui passe de 99 % (tableau II) à 42 %. On note parallèlement une forte tendance à percevoir [s] comme [ʃ]. La résistance de [s] est due cependant au maintien du pôle de stridence, qui joue un rôle important pour la perception de cette consonne. Sa valeur est moins évidente pour [f], puisqu'un [ʃ] bémolisé (dénué de pôle de stridence) est perçu comme [f] dans un contexte vocalique [- arrondi].

En revanche, on constate que le relèvement de P1 dans la séquence [u s^{i} u] ne provoque pas une modification notable de l'identification de la consonne. Ce dernier fait montre que l'on peut synthétiser un bon [s] dans tous les contextes si on prend soin de lui attribuer les indices acoustiques qu'il possède dans le contexte [-arrondi].

Cette dernière expérience fait apparaître que les effets du contexte doivent être pris en compte dans la synthèse par règles, qui doit nécessairement être de type syntagmatique. Toutefois, comme on vient de le suggérer, ces effets peuvent être simplifiés, ce qui rend possible une synthèse plus économique. Notre étude confirme, d'autre part, la validité de notre analyse acoustique pluriparamétrique, qui nous permet d'obtenir des scores de reconnaissance supérieurs à ceux de nos prédécesseurs. Cette réussite tient, en partie, à l'intégration systématique, dans le modèle de synthèse, des caractéristiques microprosodiques. D'autres recherches sont en cours, qui nous conduirons à définir la place exacte qu'occupent ces dernières dans la hiérarchie des indices qui concourrent à la perception des traits définitoires des diverses consonnes.

REFERENCES

ANDERSEN, P. , 1976, Spectral properties of German and Danish sibilants, ARIPUC, 10, pp. 29-56.

BUSH, Clara, 1964, Phonetic variation and acoustic distinctive features, Mouton, The Hague, 161 p.

CHAFCOULOFF, M. , 1976, Vingt cinq années de recherches en synthèse de la parole, Editions du C.N.R.S., 286 p.

CHAFCOULOFF, M. , DI CRISTO, A. , SEIMANDI, L. , 1976, Effets de la coarticulation sur les caractéristiques acoustiques des contoïdes fricatives du français, T.I.P.A., 3, pp. 61-113.

DANILOFF, R.G. and HAMMARBERG, R.E. , 1973, On defining coarticulation, J. of Phonetics, 1, pp. 239-248.

DELATTRE, P. C. , 1968, From acoustic cues to distinctive features, Phonetica, 18, pp. 198-230.
DELATTRE, P. C. , LIBERMAN, A. M. , COOPER, F. S. , 1964, Formant transitions and loci as acoustic correlates of place of articulation in American fricatives, Studia Linguistica, 16, pp. 104-121.
DI CRISTO, A. et CHAFCOULOFF, M. , 1978a, Notes sur l'intensité spécifique des consonnes voisées du français (à paraître), T.I.P.A., Vol. V.
DI CRISTO, A. et CHAFCOULOFF, M. , 1978b, Les caractéristiques intrinsèques de la fréquence fondamentale des consonnes voisées du français (à paraître), T.I.P.A., Vol. V.
FANT, G. , Acoustic theory of speech production, Mouton, The Hague, 328 p.
FANT, G. , LILJENCRANTS, J. , MALAC, V. , and BOROVICKOVA, B. , 1970, Perceptual evaluation of coarticulation effects, STL/QPSR, 1, pp.10-13.
HALLE, M. , 1959, The sound pattern of Russian, Mouton, The Hague, 206 p.
HARRIS, K. S., Cues for the discrimination of American English fricatives in spoken syllabes, Language and Speech, 1, pp. 1-7.
HEINZ, J. M. and STEVENS, K. N. , 1961, On the production and perception of fricative consonants, J. Acous. Soc. Am., 33, pp. 589-596.
HUGHES, G. W. and HALLE, M. , 1956, Spectral properties of fricative consonants, J. Acous. Soc. Am., 28, 303-310.
INGEMANN, F. , 1960, Formants as a voicing cue in the perception of /z/, J. Acous. Soc. Am., 32, p. 1501.
JAKOBSON, R., FANT, G. , HALLE, M. , 1952, Preliminaries to speech analysis 8th edit. , 1969, The M. I. T. Press, Cambridge.
JASSEM, W. , 1968, Acoustical description of voiceless fricatives in terms of spectral patterns, Speech Analysis and Synthesis, 1, pp. 189-206.
KACPROWSKI, R. , GUBRYNOWICZ, R. and MIKIEL, W. , 1973, Perception and intelligibility of synthetic polish fricatives, Symposium sur l'intelligibilité de la parole, Liège, oct., pp. 234-241.
LAMBERT, J. D. , 1954, The effect of frequency filtering on consonant recognition, Ph. D. thesis, Purdue University.
LEHISTE, I. and SCHOCKEY, L. , 1972, On the perception of coarticulation effects in English VCV syllabes, J. Speech Hear. Res., 15, pp. 500-506.
LIBERMAN, A. M. , INGEMANN, F. , LISKER, L. , DELATTRE, P. C. , and COOPER, F. S., 1959, Minimal rules for synthesizing speech, J. Acous. Soc. Am., 31, pp. 1490-1499.
MARTONY, J. , CEDERLUND, C. , LILJENCRANTS, J. , LINDBLOM, B. , 1962, On the analysis and synthesis of vowels and fricatives, Proceedings of the 4th Int. Congress of Phonetic Sciences, Helsinki, pp. 208-213.
NAKATA, K. , 1960, Synthesis and perception of Japanese fricative sounds, J. Rad. Res. Lab. Jap., 7, 32, pp. 319-333.
PECKELS, J. P. et ROSSI, M. , 1971, Le test de diagnostic par paires minimales, Journées d'études sur la parole, Aix-en-Provence, BD1-BH4.
ROSSI, M. , 1974, Description phonétique et phonologique du parler de Rossano, thèse de Doctorat d'Etat, Paris III.
STREVENS, P. , 1960, Spectra of fricative noise in human speech, Language and Speech, 3, pp. 32-49.
TARNOCZY, T. , 1954, Die Akustiche StruKtur der stimmlosen Engelaute, Acta Linguistica, 4, pp. 313-349.

9èmes JOURNEES D'ETUDE SUR LA PAROLE

LANNION 31 mai - 2 juin 1978

STRUCTURATION PERCEPTIVE DE DEUX INDICES ACOUSTIQUES ET PERCEPTION CATEGORIELLE.

BEECKMANS, Renaud,
Institut de Phonétique de l'Université Libre de Bruxelles,
50, avenue F. Roosevelt,
1050 Bruxelles.

.RESUME

L'ensemble des études concernant la discrimination de certains stimuli de parole (en particulier les occlusives sourdes-voisées variant suivant le délai d'établissement de voisement - V.O.T.) montrent que la discrimination obtenue par des stimuli appartenant à des classes phonémiques différentes est largement supérieure à celle obtenue pour des stimuli appartenant à une même classe phonémique.

L'interprétation de ce phénomène en terme de perception catégorielle suggère que l'effet de frontière pour le V.O.T. serait une conséquence directe de la catégorisation phonétique, processus qui supprimerait toute information discriminante autre que l'appartenance à une catégorie phonétique.

Par contre, l'interprétation plus récente en terme de discriminabilité suggère que l'effet de frontière reflète une augmentation de discriminabilité dans la région temporelle correspondant à la frontière sourd-voisé pour le V.O.T. et ce indépendamment de la catégorisation phonétique.

Une brève revue critique de ces travaux montre que l'hypothèse implicite d'un indice unique (primary cue) entraîne une confusion conceptuelle entre frontière phonémique et frontière psycho-acoustique. Dans une optique alternative en terme d'intégration perceptive de plusieurs indices acoustiques, nous proposons un modèle permettant d'extraire la fonction psycho-acoustique pour chaque indice à partir de données d'identification.

Des résultats pour des occlusives variant à la fois par le V.O.T. et le

prévoisement mettent en évidence une frontière psycho-acoustique du V.O.T. invariante par rapport aux déplacements de la frontière phonémique. Cette approche démontre la possibilité de séparer les facteurs psycho-acoustiques et phonétiques responsables du mode de perception catégoriel.

. SUMMARY

The perception of certain synthetic speech sounds, in particular stop consonants differing in voice-onset time (V.O.T.), has been found to be nearly categorical. That is, subjects can discriminate two synthetic stimuli drawn from different phonological categories better than two stimuli selected from the same phonological category.

Categorical perception has been interpreted as a specific consequence of phonetic categorization, a specialized process which strips the signal of all discriminative information other than the category label.

However, more recent results suggest further the presence of naturally determined boundaries at specific regions along the V.O.T. continuum which are distinct from phonetic processing.

A brief critical review shows the implicit hypothesis of a primary cue being responsible for a conceptual confusion between phonemic and psychoacoustic boundaries.

Based on the opposite theoretical approach viewing perception as the integration of several acoustic correlates, a quantitative model is proposed which allows to derive, from identification data, the psychoacoustic functions for each cue. Results from stop consonants which were covaried along two acoustic dimensions (positive V.O.T. and pre-voicing) points out a psychoacoustic boundary along the V.O.T. continuum which remains invariant, regardless of the shift of the phonetic boundary.

This approach provides a means of distinguishing the psychoacoustic versus phonetic factors underlying categorical perception.

L'objet de ce travail est de mettre en lumière les relations profondes existant entre les hypothèses théoriques concernant la perception des traits phonétiques et la nature de la perception catégorielle.

1. Nature perceptive du trait phonétique

On peut séparer grossièrement deux modes de pensée qui s'opposent quant à la nature perceptive des traits phonétiques, en particulier du trait de voisement.

- D'une part, les travaux entrepris aux Haskins Laboratories sur les occlusives anglaises à l'initiale, ont abouti à l'idée d'une dimension unique de voisement. Le délai séparant le début de la détente de la première impulsion quasi-périodique, appelé Voice Onset Time (VOT), constituerait l'indice essentiel (primary) du trait de voisement. Cette notion est étroitement liée à la théorie motrice dans la mesure où un certain nombre de manifestations acoustiques sont concomitantes à la commande motrice du VOT. Bien que cette théorie motrice ait été abandonnée, faute de confirmations expérimentales, la notion d'indice unique reste un postulat implicite à la base de nombreux travaux. Les études comparatives du rôle perceptif du VOT à travers plusieurs langues ont été interprétées comme confirmant, de par son caractère universel, l'unicité de cet indice: le continuum perceptif VOT serait à la base du trait de voisement pour une grande majorité de langues; seule la localisation de la frontière phonétique entre sourdes et voisées serait spécifique pour chaque langue. Un dernier argument fréquemment avancé concerne le caractère catégoriel du continuum VOT: les auditeurs ne peuvent discriminer deux stimuli différant par le délai d'établissement de voisement que dans la mesure où ils appartiennent à deux classes phonémiques différentes.
L'ensemble de ces résultats se traduit souvent par l'assimilation de cet indice au trait de voisement lui-même.

- Dans une optique toute différente, de nombreux travaux, en particulier ceux menés dans notre laboratoire depuis 1973 (Wajskop et Sweerts), se sont attachés à la notion de multiplicité des indices pertinents pour la perception du trait de voisement. Les recherches se sont portées sur deux fronts:

1. la mise en évidence d'indices pertinents outre le VOT
2. l'étude du mode de structuration de ces différents indices.

Les résultats obtenus nous ont conduit à mettre en doute la pertinence du continuum VOT: dans le cas des occlusives voisées en français, la détente est précédée de vibrations laryngées (prévoisement) ce qui correspond d'après Lisker et Abramson (1964) à une valeur de VOT négative. Une expérience d'identification à partir de stimuli variant à la fois suivant le VOT positif et le prévoisement (Beeckmans et Serniclaes, 1975) a montré que ces deux indices jouaient un rôle perceptif autonome et ne constituaient en aucun cas, deux régions d'un même continuum perceptif. Dès lors, la frontière phonémique pour le VOT varie largement suivant la valeur du prévoisement. Avant d'envisager l'implication de ce résultat pour le mode de perception catégorielle, il nous paraît utile de rappeler les principaux résultats concernant ce type de perception.

2. Perception catégorielle

L'évolution chronologique des différentes hypothèses concernant le statut de la perception catégorielle a engendré une situation circulaire suite à un fondement théorique implicite inadéquat.

De très nombreux travaux depuis Liberman (1957) ont montré que des consonnes synthétiques variant suivant un paramètre physique sont hautement discriminables lorsque les consonnes sont rangées dans des classes phonétiques différentes, et à l'inverse, peu discriminables lorsqu'elles sont rangées dans la même classe phonétique. En ce qui concerne les traits de lieu d'articulation et de voisement, les résultats sont très proches de ceux qu'on obtiendrait en faisant l'hypothèse extrême que l'appartenance à l'une ou l'autre classe constitue le seul critère utilisable par le sujet. Ce mode de perception est incompatible a priori avec les résultats classiques de psycho-physique dont il ressort que les possibilités de discrimination sont nettement supérieures aux possibilités de jugement absolu (Pollack, 1952). Les premiers travaux menés aux Haskins Laboratories (Liberman, 1961) ont posé le problème du caractère inné ou acquis de la perception catégorielle en comparant les courbes de discrimination de consonnes synthétiques d'une part et de stimuli de non-parole (inversion spectrale des consonnes) d'autre part. Les résultats pour les stimuli de non-parole étant non catégoriels et très inférieurs par rapport à ceux obtenus pour les stimuli phonétiques, les auteurs concluent que les sujets acquièrent une meilleure discrimination de part et d'autre de la frontière phonémique. Les travaux ultérieurs confirment la spécificité de la perception catégorielle aux consonnes par rapport à ce genre de stimuli de non-parole et également par rapport aux voyelles. La perception ca-

tégorielle devient alors un argument essentiel de la théorie motrice (Liberman et al, 1962) : les discontinuités articulatoires se retrouveraient au niveau perceptif. A partir de ces mêmes résultats, Lane (1965) propose une théorie alternative: la différence critique entre les continua de parole et de non-parole résiderait dans le fait que seul le continuum de parole est divisé en classes. Il montre que l'on peut obtenir une perception légèrement catégorielle pour un continuum visuel, les sujets ayant été entraînés au préalable à classer les différents stimuli en deux classes. Les résultats ne sont pas cependant tout à fait convaincants; c'est le cas également pour des reduplications ultérieures (Parks et al, 1969; Liberman et al, 1965) de sorte qu'on ne peut, contrairement au cas de la théorie motrice, écarter définitivement l'hypothèse proposée par Lane.

Des études plus récentes ont fourni des évidences expérimentales qui ont modifié radicalement l'interprétation du mode de perception catégorielle. Une première série d'expériences concerne la perception des stimuli phonétiques chez les très jeunes enfants. Eimas (1971) a montré que des enfants âgés de un à trois mois seulement discriminaient déjà des stimuli synthétiques variant suivant la dimension du VOT de manière relativement catégorielle. Des expériences analogues avec des enfants élevés dans des contextes linguistiques différents, espagnol (Lasky et al, 1975) et kikuyu (Streeter, 1976) mettent en lumière une perception catégorielle autour de frontières identiques et ce, bien que ces frontières ne soient pas phonétiquement pertinentes pour ces langues. En conséquence, la discrimination catégorielle observée semble indépendante du contexte linguistique et dépendrait uniquement des attributs psycho-acoustiques des stimuli employés.

Une seconde série de résultats qui mettent en évidence une perception nettement catégorielle pour certains types de continua de non-parole ont permis de préciser la nature psycho-acoustique de la frontière qui divise le continuum en deux catégories perceptives différentes. Par exemple, des stimuli constitués de deux sons séparés par un intervalle de silence variable seront perçus comme un ou deux sons de part et d'autre d'une valeur critique de cet intervalle. Cette frontière "naturelle" induirait à son voisinage une meilleure discrimination indépendante par conséquent d'une catégorisation apprise. Partant de cette idée, Miller et al (1976) et Pisoni (1977) ont suggéré qu'en ce qui concerne le VOT, l'existence d'une frontière de ce type pourrait expliquer le caractère catégoriel de la perception du trait de voisement. Les résultats de ces études montrent une perception nettement catégorielle pour des stimuli constitués de deux sons différant dans le temps. La variation de délai induirait une perception de nature différente quant à l'ordre des deux composants (avant, pen-

dant, après) comparable à l'effet du VOT pour les consonnes en position initiale. Le mode catégoriel de la perception de la parole ne serait pas une conséquence de la catégorisation phonémique mais, à l'inverse, se serait développé autour de propriétés non spécialisées existantes du système perceptif. Cette théorie présente au moins deux aspects positifs :

1. la contradiction apparente entre perception catégorielle et résultats classiques de psycho-physique tombe dès lors que le continuum physique recouvre en fait deux sensations auditives différentes.
2. l'hypothèse d'une organisation de la communication parlée à partir de discontinuités perceptives constitue une alternative séduisante à la théorie motrice et, en tout cas, remet en cause les statuts respectifs production-perception.

Néanmoins, telle quelle, cette théorie semble insuffisante pour expliquer les importantes possibilités d'adaptation à un entourage linguistique donné.

Il ressort de cette brève synthèse que la perception catégorielle qui a été pendant quinze ans l'argument essentiel pour l'hypothèse de mécanismes hautement spécialisés pour la perception de la parole est actuellement l'argument essentiel d'une hypothèse opposée en termes de processus psycho-physiques très généraux. La raison évidente de cette situation réside dans l'hypothèse implicite de l'indice unique: dans cette optique, en effet, les frontières phonétique et psycho-acoustique confondues dans les faits (neutralisation des autres indices) sont abusivement confondues conceptuellement.

Par contre, si l'on considère la perception d'un trait phonétique comme l'intégration de différents corrélats acoustiques, la frontière psycho-acoustique pour un indice, si elle existe, est tout-à-fait différente de la frontière phonétique puisque cette dernière dépend également des autres indices. Comme nous l'avons dit en début de travail, un déplacement important de la frontière phonémique a été mis en lumière par une expérience d'identification manipulant simultanément deux indices de voisement (Beeckmans et Serniclaes, 1975). Dans ces conditions, il nous a paru très utile de tester l'existence à partir de ces résultats, d'une frontière psycho-acoustique pour le VOT.

3. Relations entre l'intégration phonétique et le mode de perception catégorielle

La covariation de plusieurs corrélats étant nécessaire sur le plan expérimental pour différencier les niveaux psycho-acoustique et phonétique, la même distinction doit se retrouver conjointement en ce qui concerne le

modèle théorique et le traitement des résultats. De nombreuses formalisations du processus perceptif considéré comme l'intégration de plusieurs indices ont été développées avec succès dans le domaine de la mesure en psycho-physique. Il est remarquable que ces travaux aient connu une évolution strictement identique à celle qui ressort des études sur la perception catégorielle: dans un premier temps, la recherche d'une loi psycho-physique à partir de résultats limités à une seule variable a abouti à des résultats contradictoires tels la controverse entre fonction de puissance et fonction logarithmique (Treisman, 1962).
Par contre, la prise en considération de deux variables et leur intégration a permis d'isoler les échelles psycho-physiques par rapport aux autres fonctions du processus perceptif. (Pour une revue détaillée voir Anderson, 1974). Une formalisation simple à partir de ces méthodes rend ce fait évident: la réponse à un stimulus variant suivant une dimension physique unique (intensité, durée...) peut être considérée comme la composée de deux fonctions:

1. la fonction psycho-physique

$$\Psi = H\ (p)$$

reliant à la mesure physique p du stimulus, une mesure de l'effet neural Ψ de ce stimulus.

2. la fonction réponse

$$R = J\ (\Psi)$$

associant à l'effet neural, la réponse observable R.

La fonction composée R = HoJ (p) étant seule connue, on ne peut évidemment spécifier la nature de la fonction psycho-physique sans faire d'hypothèse a priori concernant la nature de la fonction R.

Par contre, pour deux variables, les relations deviennent :

$$\Psi_1 = H_1\ (p_1) \qquad \text{et} \qquad \Psi_2 = H_2\ (p_2)$$

$$R = J\ (\Psi_1, \Psi_2)$$

Exprimée sous cette forme, chacune des deux échelles psycho-physiques Ψ_1 et Ψ_2 relatives à chacune des deux variables physiques p_1 et p_2 reste invariante par rapport aux valeurs de l'autre variable, ce qui formalise la distinction entre le niveau psycho-physique et le niveau d'intégration perceptive. Cette seule hypothèse, qui peut être testée par l'adéquation du modèle aux données, permet dans de nombreux cas la détermination univoque de chacune

des fonctions (Anderson, 1970). L'application de ce modèle à nos résultats d'identification présentait une difficulté de par la nature dichotomique des réponses (sourd-voisé). Nous avons développé une méthode itérative permettant de déterminer à partir de ce genre de données les échelles psycho-physiques correspondant à chaque paramètre physique (Beeckmans, 1977). Par cette méthode, la validité des échelles psycho-physiques ainsi que le mode d'intégration des variables peuvent être testés directement par un procédé dérivé de la méthode Probit (Finney, 1971). Les résultats de cette étude ont montré pour chacun des quatre sujets expérimentés:

- la validité de l'hypothèse de base du modèle: l'échelle psycho-physique obtenue pour le VOT est indépendante des valeurs de prévoisement;
- une non-linéarité statistiquement significative de cette échelle présentant une zone de meilleure précision pour les valeurs situées entre 8 et 12 ms.

Ces résultats fournissent une confirmation directe évidente de la théorie psycho-physique, mais en outre, il en ressort clairement que la frontière psycho-acoustique située aux environs de 10 ms pour le VOT, quelle que soit la valeur de l'autre indice, est donc différente de la frontière phonétique qui elle varie largement (0 → 20 ms au moins) en fonction de la valeur du prévoisement.

Des expériences de discrimination portant sur des stimuli de même nature (covariation du VOT et du prévoisement) sont actuellement en cours qui devraient permettre, par comparaison avec ces premiers résultats, de départager les effets phonétiques des effets psycho-acoustiques pour la perception catégorielle.

Références

ANDERSON, N.H., 1970, Functional measurement and psychophysical judgment. Psychological Review. 77, pp. 153-170.

ANDERSON, N.H., 1974, Algebraic models in perception. In E.C. Carterette and M.P. Friedman (Eds.), - Handbook of perception, Volume 2, pp. 215-298.

BEECKMANS, R., 1977, Description et application d'une méthode itérative de transformation d'échelle permettant de tester différentes hypothèses d'intégration de deux indices, à partir de données d'identification. - Rapport d'Activités de l'Institut de Phonétique - U.L.B. 11/1, pp. 69-82.

BEECKMANS, R., 1978, Détermination de courbes psychophysiques pour deux corrélats acoustiques du trait de voisement. - Rapport d'Acti-

vités de l'Institut de Phonétique - U.L.B. (à paraître).

BEECKMANS, R. and SERNICLAES, W., 1975, Positive and negative Voice Onset Time: two independent cues in voicing perception. - Proc. of the 8th Int. Cong. of Phon. Sci. (à paraître).

EIMAS, P.D., SIQUELAND, E.R., JUSCZYK, P. and VIGORITO, J., 1971, Speech Perception in infants. - Science, 171, pp. 303-306.

FINNEY, D.J., 1971, Probit Analysis - Cambridge: University Press.

LANE, H., 1965, The Motor Theory of Speech Perception: A Critical Review. - Psychological Review. 72, pp. 275-309.

LASKY, R.E., SYRDAL-LASKY, A., and KLEIN, R.E., 1975, VOT discrimination by four to six and a half month old infants from Spanish environments. - J. Exp. Child Psychol. - 20, pp. 213-225.

LIBERMAN, A.M., 1957, Some results of research on speech perception. - J.A.S.A. 29, pp. 117-129.

LIBERMAN, A.M., HARRIS, K.S., EIMAS, P.D., LISKER, L., and BASTIAN, J., 1961, An effect of learning on speech perception: the discrimination of duration of silence with and without phonemic significance. - Language and Speech. 4, pp. 175-195.

LIBERMAN, A.M., COOPER, F.S., HARRIS, K.S., and MAC NEILAGE, P.F., 1962, A motor theory of speech perception. - Paper read at the Speech Communication Seminar, Speech Transmission Laboratory, Royal Institute of Technology, Stockholm.

LIBERMAN, A.M., STUDDERT-KENNEDY, M., HARRIS, K.S. and COPPER, F. S., 1965, A reply to "Identification and discrimination functions for a visual continuum and their relation to the motor theory of speech perception" by Cross, Lane, and Sheppard. SR-3. New York: Haskins Laboratories.

LISKER, L. and ABRAMSON, A.S., 1964, Cross-language study of voicing in initial stops: Acoustical measurements. - Word. 20, pp.384-422.

MILLER, J.D., WIER, C.C., PASTORE, R., KELLY, W.J., and DOOLING, R.J., 1976, Discrimination and labeling of noise-buzz sequences with varying noise-lead times: An example of categorical perception. - J.A.S.A.. 60, pp. 410-417.

PARKS, T., WALL, C., and BASTIAN, J., 1969, Intercategory and intracategory discrimination for one visual continuum. - Journal of Experimental Psychology. 81, pp. 241-245.

PISONI, D.B., 1977, Identification and discrimination of the relative onset time of two component tones: Implications for voicing perception in stops. - J.A.S.A. 61, pp. 1352-1361.

POLLACK, I., 1952, The information of elementary multidimensional auditory displays. - J.A.S.A. 26, pp. 155-158.

STREETER, L.A., 1976, Language perception of 2- month old infants shows effects of both innate mechanisms and experience. - Nature. 259, pp. 39-41.

TREISMAN, M., 1964, Sensory seating and the psychophysical law. - Quartely Journal of Experimental Psychology. 16, pp. 11-22.

WAJSKOP, M. and SWEERTS, J., 1973, Voicing cues in oral stop consonants. - Journal of Phonetics. 1, pp. 121-130.

THEME 2

DESCRIPTION ARTICULATOIRE ET ACOUSTIQUE DES INDICES ET DES TRAITS

9èmes JOURNEES D'ETUDE SUR LA PAROLE

LANNION 31 mai - 2 juin 1978

INDICES ET TRAITS ACOUSTIQUES DANS UN SYSTEME DE RECONNAISSANCE DE LA PAROLE CONTINUE : QUELQUES RESULTATS

J . CAELEN et G. PERENNOU

Laboratoire C.E.R.F.I.A.

118, route de Narbonne - 31077 TOULOUSE CEDEX

.RESUME

Dans un système de reconnaissance de la parole continue toutes les informations acoustiques doivent être utilisées si l'on ne veut pas alourdir les étages linguistiques.

Par ailleurs, il apparait difficile d'opérer en catégorisations successives: phonèmes, syllabes, mots, ou à l'inverse : hypothèse, phrases, mots, syllabes, phonèmes. En effet, les indices acoustiques sont parfois trop fragiles pour autoriser la construction d'unités linguistiques et d'autre part, la méthode descendante qui procède par vérification (et de ce fait, demande des indices moins précis) est inadéquate à partir d'un facteur de branchement, au-delà de quelques dizaines.

On décrit ici un processus de reconnaissance de la parole continue, au niveau acoustique, basée sur l'utilisation d'indices. On montre comment il est possible de transformer ces informations en traits acoustiques. Des résultats sont fournis qui montrent la corrélation entre les traits acoustiques et les catégories phonétiques classiques.

La stratégie générale prévoit que les traits acoustiques peuvent être acquis en cas de vraisemblance suffisante sinon, les segments restent en attente d'une demande de vérification.

SUMMARY

INDEXES AND ACOUSTICAL FEATURES
IN A SYSTEM OF CONTINUOUS SPEECH RECOGNITION :
A FEW RESULTS

J. CAELEN et G. PERENNOU

Laboratoire C.E.R.F.I.A.
118, route de Narbonne - 31077 TOULOUSE CEDEX

In a connected speech recognition system, each acoustical information is to be taken into account in order to avoid a sophistication of linguistical stages. On the other hand, it seems ineffective to process by successive categorization : phoneme, syllabe, word or conversely : hypothesis, sentence, word, syllabe, phonem.

Sometimes, acoustical features are indeed too weak to give linguistics units and moreover, the top-to-bottom method proceeding by matching (needing thereby less accurate features) is inefficient when the branching factor exceeds a few tens.

A process of continuous speech recognition on the acoustical stage, founded on the use of indexes is described here. We explain how to convert these informations into acoustical categories. Results are given pointing out the correlation between acoustical and classical phonetical categories.

In the ordinary strategy acoustical features may be given if there is sufficient likelihood, otherwise, segments are waiting for a matching request.

INTRODUCTION

Le signal vocal est analysé à partir du modèle d'oreille (CAELEN, J., 1974) Les paramètres obtenus permettent de dégager clairement la structure formantique ainsi que divers autres indices. Cela permet d'envisager la reconnaissance de la parole à partir du modèle d'oreille (CAELEN, J., PERENNOU G., 1976 et CAELEN,J., EL JAI, M.C., PERENNOU, G., 1977).

Nous nous proposons de préciser ici les indices et les traits acoustiques utilisés ainsi que l'organisation générale de l'étage de reconnaissance syllabique de notre système.

2. ORGANISATION GENERALE DU SYSTEME DE RECONNAISSANCE SYLLABIQUE

Le modèle d'oreille transforme le signal vocal s(t) en une suite de données spectrales relatives à des intervalles successifs de 8,5 ms. Au temps t_n ces données spectrales seront :

E_{on} = fréquence fondamentale ou 0,

W_{kn} = énergie en dB au-dessus du bruit moyen, du $k^{ème}$ canal ($1<k<24$).

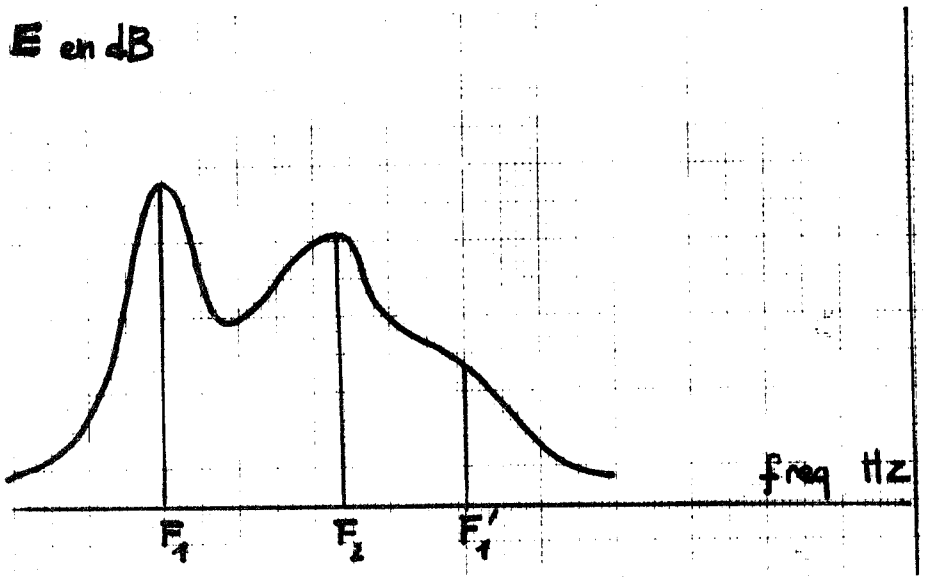


Fig. 1
Détermination des formants (F_1,F_2) et des fréquences masquées (F'_1)
Determination of formants (F_1,F_2) and masked frequencies (F'_1).

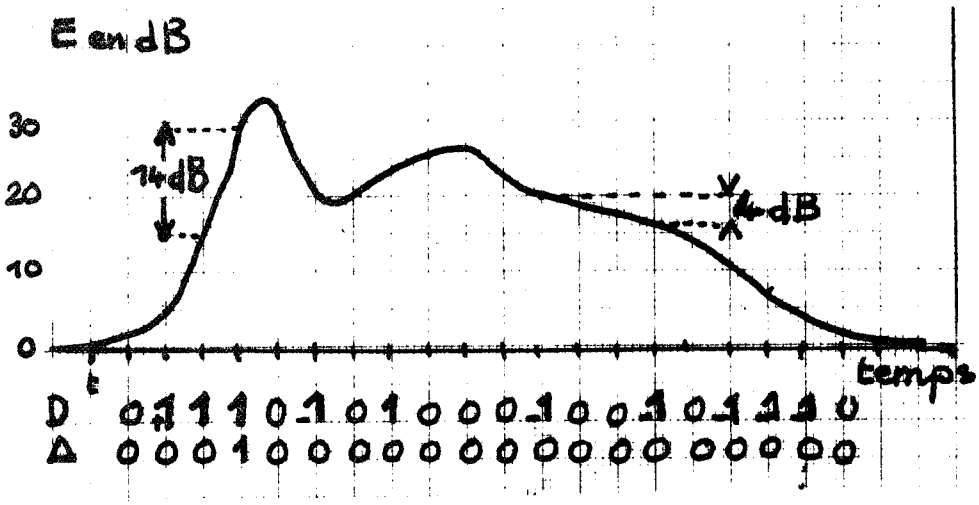


Fig. 2
Indices D et Δ
Indexes D and Δ

Nous poserons $S_n = (W_{kn})_{1<k<n}$.

Paramètres spectraux instantanés.

D'autres paramètres sont déduits des précédents. Ce sont :

$E_n = \Sigma W_{kn}$ qui rend compte de l'énergie,

F_{kn} = fréquence de la $k^{ième}$ crête du spectre continu résultant d'une interpolation polynomiale,

ν_n = nombre de crêtes

aux formants ainsi détectés s'ajoutent ceux qui sont masqués. Pour les caractériser au mieux sans trop de calcul nous opérons par les maxima de concavité négative, ce qui donne ν'_n fréquences $F'_{\ell n}$ $(1 \leqslant \ell \leqslant \nu'_n)$; (voir figure 1).

WF_{jn}(resp. WF'_{jn}) = énergie à la $j^{ème}$ fréquence F_{jn}(resp.F'_{jn}) ;

$GA_n = (W_{1n}+W_{2n}+W_{3n})/(W_{22n}+W_{23n}+W_{24n})$ mesure le rapport entre l'énergie dans le grave (G) et dans l'aigu (A).

Indices spectraux instantanés

Ces indices sont en relation avec un trait phonétique. Donnons-en deux :

Nas_n = SI (230 Hz$<F_{1n}<$300 Hz) ALORS 0 SINON 1 ;

$Fric_n$ = SI ($\exists$ k)(((F_{kn} > 3200 Hz) OU(2000 Hz $\leqslant F_{kn} \leqslant$ 3200 Hz ET WF_k dominante) OU (ν_n = 1) ET F_1 >1500 Hz)) ALORS 1 SINON 0 ;

Le premier est un indicateur de formant nasal. Bien entendu, il peut aussi indiquer une occlusive voisée et certains épisodes de /ℓ/.

Le second est un indicateur de friction. Il peut être 1 également pendant l'explosion d'une occlusive ou par suite de bruits parasites.

Indices et paramètres d'évolution.

Ils traduisent l'évolution de t_{n-1} à t_n ou à plus long terme

ΔE_n = SI E_i-E_{i-1} > 12 dB ALORS 1 SINON

SI $E_i-E_{i-1} \leqslant$ -12 dB ALORS -1 SINON 0 ;

DE_n = SI E_i-E_s > 3 dB ALORS 1 SINON

SI E_i-E_s < -3 dB ALORS -1 SINON 0 ;

avec E_s = Energie au dernier extremum de la courbe d'énergie ou celle au temps t_m tel que ΔE_n ou $DE_n \neq 0$.

La figure 2 illustre le rôle de ces deux indices dans la détection des variations significatives d'énergie du point de vue de la segmentation.

$$\bar{w}_n = (\sum_{k=1}^{24} w_{kn})/24,$$

$$\Delta sp_n = \sum_{k=1}^{24} |(w_{kn}-\bar{w})-(w_{kn-1}-\bar{w}_{n-1})|$$ cet indice mesure la variation énergétique du spectre ; il est semblable à ceux utilisés par MERCIER,G., 1976, MLOUKA, M. et LIENARD, J.S., en 1974.

$$\Delta f_n = \sum_{\ell=1}^{\nu_n} \log_2(F_{\ell n}/F_{\ell n-1})$$

Segmentation.

Elle est effectuée à partir des indices et paramètres ΔE_n, DE_n, Δsp_n, Δf_n, variations de Fric et de Nas. Ceci n'étant pas l'objet de l'article, nous ne mentionnerons ce traitement que pour le situer dans l'ensemble du traitement. Indiquons seulement qu'en général il y a plus de segments que de phonèmes. Les plosives en particulier se découpent souvent en un segment occlusif et un segment explosif.

Il peut arriver qu'une frontière soit omise, notamment en présence de semi-voyelles.

Paramètres et indices segmentaux.

Ces indices sont évalués pour l'ensemble d'un segment S_m

$Ifric_m$ = SI ($Fric_n$ = 0 sur S_m) ALORS 0 SINON
SI ($Fric_n$ = 1 sur moins de la moitié de S_m) ALORS 1
SINON 2 ;

$INas_m$ = SI (Nas_n = 0 sur moins des 3/4 de S_m) ALORS 0
SINON 1 ;

$Imax_m = E_n$ max sur S_m

$Imin_m = E_n$ min sur S_m

νmin_m (resp. νmax_m) : nombre de maxima (resp.minima)sur l'énergie filtrée avec un passe-bas de 200 ms de constante de temps.

Pour ces deux derniers indices on peut noter une analogie avec la mémoire acoustique à court terme de l'humain.

T_m = durée du segment en nombre d'intervalles de 8,5 ms ;
TV_m = durée du voisement sur l'intervalle
IC_m = SI $DE_n > 0$ à l'origine du segment ET
$DE_n < 0$ à la fin du segment ALORS 1
SINON 0 ;

ID_m = SI $DE_n \leqslant 0$ sur tout le segment ALORS 1
SINON 0 ;

La figure 3 donne l'organisation d'ensemble de l'étage assurant la segmentation et la recherche des indices. Il débouche sur la recherche de traits acoustiques et la reconnaissance des syllabes. Il est à noter que cet étage n'est nullement figé et que certains indices peuvent facilement être modifiés et d'autres rajoutés ou supprimés.

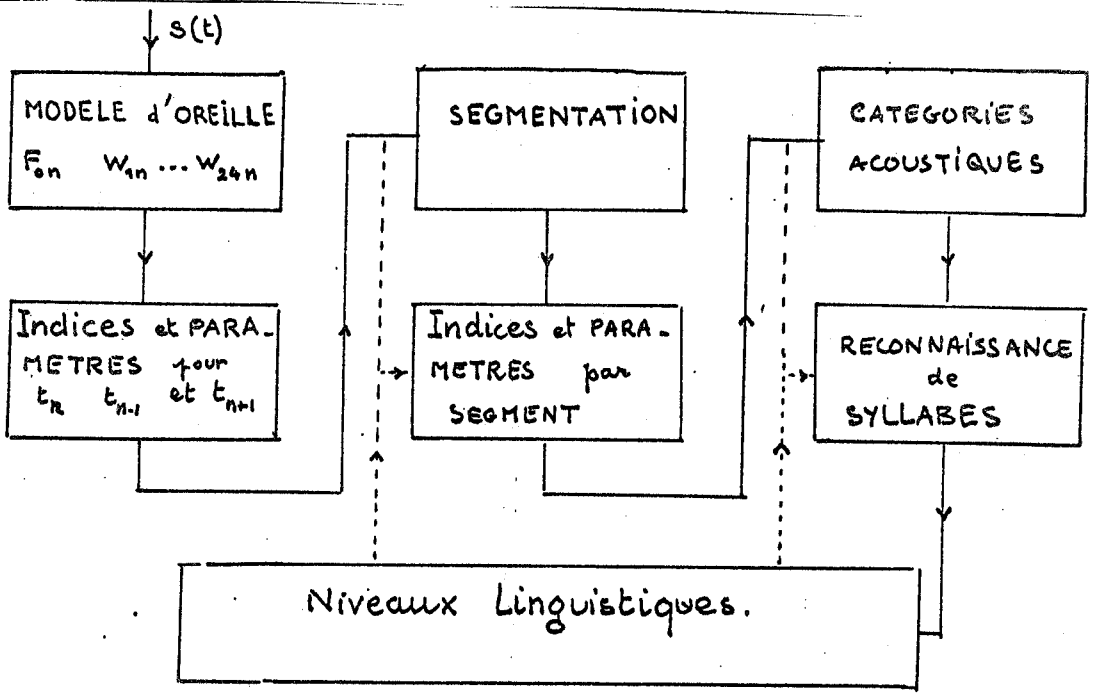


Fig. 3
Schéma de principe de l'étage acoustique
Schematic representation of the acoustical stage

ATTRIBUTION DES TRAITS ACOUSTIQUES

Pour l'attribution d'une étiquette recouvrant un ou plusieurs traits acoustiques on utilise la table ci-après. Chaque ligne correspond à une condition qui implique à des degrés divers l'étiquette donnée en colonne.

-2 : rend l'étiquette invraisemblable
-1 : rend l'étiquette peu vraisemblable
0 : condition indifférente pour l'étiquette
1 : rend l'étiquette plausible
2 : rend l'étiquette vraisemblable

Les conditions ne sont pas exhaustives et certaines ont trait au contexte. La signification des étiquettes est évidente. Indiquons cependant que V signifie voisé (dans OCCV et FRICV) FLU : fluide, et TRANS : transition.

Il convient de ne pas établir des relations trop précises avec les phonèmes à ce niveau . Ainsi un /n/ de liaison peut fort bien avoir un bilan négatif pour l'étiquette NAS et positif pour FLU.

Les bilans par colonnes sont comparés à des seuils et les étiquettes suffisamment assurées sont attribuées. De même, d'autres sont refusées.

Pour les autres, il faudra attendre les hypothèses de l'analyse descendante afin de procéder par vérification.

Conditions liées au segment	OCCS	CON	VOY	FRICS	EXPL	TRANS	OCCV	FLU	NAS	FRICV
Imin ⩽ 3	2		-2	-2			-1	-2	-2	-2
3 < Imin ⩽ 10	1		-2	-1				-1	-1	-2
Imin ⩾ 10	-2									2
T<9				-2		1				
T⩾9					-2	-2				
T⩽4				-2	1	1				-1
T⩾4					-2					
Ifric = 2			-2	2				-1	-1	1
Ifric = 1			-1	1						1
Ifric = 0				-2						
GA > 200	-2		-2	-2	-1		1	-1	1	-1
175 < GA ⩽ 200	-1		-1	-2			1		1	
110 < GA ⩽ 175						-2				
GA ⩽ 110							-2		-2	
ν_{max} = 1 ET ν_{min} = 0	-2	-1	1				-2		-1	
νmin = 1 ET νmax = 0	1	1	-1				1	1	1	
Inas = 1	-2		-1	-2	-2		1		1	
Inas = 0							-2		-1	
TV>0	-2			-2	-2		1			
TV = 0							-1			
IC = 1	-1	-1	1		1		-1			
ID = 1	1	1	-1		-2		1	1	1	
TV ⩾ 3/4 T	-2			-2						
Imax - Imin > 25 ET TV = 0	1		-1							

Conditions liées au contexte										
Segment précédent = NAS				-1	-1		-1	-1		-1
Segment précédent = OCC ou OCCV					1	1			-1	
Segment précédent ≠ OCC ou OCCV					-1					

Tableau pour l'étiquetage en catégories acoustiques

Voici quelques résultats précisant la corrélation des étiquettes acoustiques avec les catégories phonétiques correspondantes, ceci pour un locuteur sur 19 phrases. + ETIQ (resp. - ETIQ) attribue (resp. refuse) l'étiquette ETIQ.

résultats / indice de		succès	échec	commentaires
Friction	+	94 %	6 %	trop de frictions affirmées
	-	100 %	0 %	
Nasalité	+	77 %	23 %	trop de nasalités affirmées
	-	86 %	14 %	trop peu " "
Voisement	+	100 %	0 %	trop de voisements affirmés
	-	98 %	2 %	trop peu " "

Etiquette	réussite %	commentaires
OCC	100 %	
CONS	91 %	
VOY	97 %	confusionsavec /r/
EXPL	89 %	
NAS	96 %	
FLU = LIQ	60 %	faible !
OCCV	91 %	
FRIC	100 %	
FRICV	—	

On constate que l'étiquette FLU recouvre mal les liquides. L'étude de ces dernières et des semi-voyelles nécessite en effet une stratégie plus complexe dans le suivi des formants et la caractérisation de l'évolution de l'énergie.

Nous renvoyons à WEINSTEIN, C.J. et autres (1975) pour des exemples de traitement de ce genre.

CONCLUSION

L'étage acoustique d'un système de reconnaissance de la parole continue semble devoir être structuré de manière à permettre l'utilisation des informations disponibles. Elles proviennent aussi bien du signal vocal que de contraintes permanentes : structure des phrases, des mots et des syllabes ainsi que des hypothèses qu'autorise la pragmatique et que génère le système d'analyse.

On constate néanmoins que toute l'information acoustique disponible doit être utilisée pour obtenir de bonnes performances. C'est l'objectif que nous poursuivons en utilisant l'analyse spectrale provenant d'un modèle d'oreille, ce qui semble donner des paramètres de bonne qualité, et en élaborant une stratégie souple pour l'utilisation des informations acoustiques.

REFERENCES

CAELEN, J., 1974 : Un modèle mathématique de cochlée et son application à l'analyse de la parole. Thèse Docteur-Ingénieur - Toulouse.

CAELEN, J., PERENNOU, G., 1976 : Un modèle d'oreille appliqué à l'analyse de la parole - 7èmes Journées d'Etude sur la Parole - Nancy.

CAELEN, J., EL JAI, M.C., PERENNOU, G., 1977 : Méthodes structurelles en reconnaissance de la parole au niveau acoustique inspirées par un modèle d'oreille. Colloque GRETSI - Nice.

MERCIER, G., 1976 : Analyse phonétique. Recherches acoustiques CNET Lannion. Vol. III.

MLOUKA, M., LIENARD, J.S., 1974 : Reconnaissance par mots basée sur les transitions phonétiques. 5èmes Journées d'Etude sur la Parole - Orsay.

ROSSI, M., 1975 : Les contraintes phonologiques dans un système de reconnaissance de la parole. 6èmes Journées d'Etude sur la Parole - Toulouse.

ROSSI, M., 1971 : Les traits et indices. 3èmes Journées d'Etude sur la Parole. Lannion.

WEINSTEIN, CJ., Mc CANDLESS, SS., MONDSHEIM, LF. et ZUE, VW., 1975 : A system for acoustic-phonetic analysis of continuous speech. IEEE (Ac, Sp and Signal Proc) n° 1 vol. ASSP-23

9èmes JOURNEES D'ETUDE SUR LA PAROLE

LANNION 31 mai - 2 juin 1978

INDICES DE DETECTION DE FORMANTS SUR ANALYSE SPECTRALE PAR CANAUX

ROSSI Mario, Institut de Phonétique, Aix-en-Provence, LA.CNRS, 261.
LE CORRE Christian, CNET, Lannion
MERCIER Guy, CNET, Lannion

.**RESUME**

On définit des indices qui permettent de détecter les formants à partir des maxima d'énergie. La présence d'un premier indice relatif à la valeur de l'énergie dans les canaux (4 à 8) autorise l'effacement du pôle superflu qui apparaît notamment sur les voyelles antérieures (/y/, /i/). La détection de F2 et F3 est consécutive à l'effacement de ce pôle et au cadrage des Max. 3 et 4. Un deuxième indice, fondé sur une évaluation de la largeur de bande des Max. permet de restituer les formants non détectés, lorsque le pôle d'énergie correspondant n'apparaît pas. C'est le cas, par exemple, pour /a/, /o/ et /y/, où respectivement F1/F2 et F2/F3 sont très voisins et confondus dans l'analyse par canaux. On calcule un coefficient qui constitue l'indice de largeur de bande. On propose un algorithme qui, sur la base des indices ci-dessus, conduit au cadrage des Max. d'énergie et à la détection des formants.

CUES FOR FORMANT DETECTION IN VOCODER SPECTRAL ANALYSIS

ROSSI Mario, Institut de Phonétique, Aix-en-Provence, LA.CNRS, 261.
LE CORRE Christian, CNET, Lannion.
MERCIER Guy, CNET, Lannion.

. SUMMARY

Cues are defined which make it possible to detect formants from the energy poles. The presence of a first cue relative to the energy level in the channels (4 to 8) makes it possible to wipe out the spurious pole (fig. 1, ALGO I) characteristic of the front vowels (/y/, /i/). The detection of F2 and F3 follows the removal of this pole and the centring on poles 3 and 4 (fig. 2, ALGO II).

A second cue based on the band-width of the poles makes it possible to reconstitute undetected formants when the corresponding energy pole is missing. This is the case, for example, for /a/, /o/ and /y/ where, respectively F1/F2 and F2/F3 are very close and not distinguished in the Vocoder analysis. A coefficient is calculated which constitutes the cue for bandwidth. An algorithm is proposed which, on the basis of the abovementioned cues makes it possible to centre on the energy poles and detect the formants correctly. (figs 1 & 2). The results obtained (fig. 3) show that the error-rate drops, with the application of this model, from 49 to 7 %. The algorithm is subsequently modified in order to correct this residual error which is essentially due to a faulty identification of F3.

INTRODUCTION

L'analyse spectrale par canaux fournit des maximums d'énergie qui ne représentent pas toujours les formants des voyelles. Ainsi les maximums détectés à partir d'un vocodeur à 14 canaux se trouvent en correspondance avec les formants dans 51 % des cas seulement. Plusieurs procédures de poursuite de formants ont été élaborées ; les principales sont : l'analyse par synthèse appliquée à la sortie d'un banc de filtres (Bell, Fujisaki, Heinz, Stevens and House, 1961 ; Olive, 1971), l'application de contraintes sur la fréquence et l'amplitude des formants aux résultats de la transformée en z de Chirp (Rabiner, Schafer and Rader, 1969 ; Schafer and Rabiner, 1970), les techniques de plus grande vraisemblance (Itakura and Saito, 1970) et de prédiction linéaire (Atal and Hanauer). Certaines de ces techniques sont efficaces bien entendu, mais elles sont également longues et coûteuses. Or, dans l'immédiat, nous devons utiliser au mieux les données spectrales d'un vocodeur à canaux et par conséquent mettre en oeuvre une méthode simple et rapide.

La discordance entre les maximums et les formants a deux causes essentielles :

1) Sur certaines voyelles antérieures, telles que [i] et [y], apparaît, aux alentours de 1000 Hz, un pic d'énergie qui n'est pas prévu par la fonction de transfert du conduit vocal (Fant, 1960).

Ex. : Echantillon 49 sur [i] de ingratitude

Canaux :	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Energie:	13	9	5	6	4	8	11	11	11	7	6	9	10	6
Maximums	!			!				!					!	
détectés:	490			943				2072					3516	

Si on conserve ce maximum, la voyelle [i] pourra être reconnue comme [u] ou [o].

2) Sur certaines voyelles comme [a], [ɔ], [y] et [i], les pics d'énergie qui correspondent à F2 ou à F3 peuvent ne pas exister. Cela se produit par exemple sur la voyelle [ə] où F1 et F2 sont très voisins, sur [y] où F2 et F3 sont parfois confondus et sur [i] où F3, selon le contexte, est proche de F2 ou de F4.

Ex. : Echantillon 12 de la voyelle [a] dans tape.

Canaux :	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Energie	11	12	13	13	11	11	8	7	9	8	5	8	7	5
Maximums			!						!			!		
détectés :			1043						2463			3502		

Cela se produit également sur des voyelles où, pour une raison ou une autre, ces formants semblent avoir été amortis.

Ex. : Echantillon 6, sur la voyelle [i] de quiconque

Canaux :	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Energie :	13	9	5	4	4	3	4	7	7	7	7	7	7	4
Maximums		!									!			
détectés :		686									2768			

Dans d'autres cas, sur les voyelles nasales par exemple, F2 ou F3 ont été effacés par une antirésonance :

Ex. : Echantillon 28, sur la voyelle [ɔ̃] de quiconque

Canaux	:	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Energie	:	12	12	12	9	9	6	5	5	3	3	5	9	11	5
Maximums					!									!	
détectés	:				1009									3351	

Il est essentiel de pouvoir identifier les deux premiers types de fusion et de les distinguer des effets de l'antirésonance. Dans le premier cas, en effet, la fusion peut être un indice d'ouverture ou de labialité de la voyelle, dans le second cas un indice de compacité de la consonne adjacente, dans le troisième enfin un indice de nasalité.

Pour interpréter correctement ces indices et supprimer les erreurs introduites par ces anomalies apparentes dans la reconnaissance des voyelles, on doit effacer les maximums superflus et restaurer ceux qui manquent. Cette opération est nécessaire pour la reconnaissance des consonnes si celle-ci est fondée en particulier sur l'évolution des formants.

ALGORITHME DE CADRAGE DES FORMANTS

L'algorithme que nous proposons pour le cadrage des formants utilise 1) la valeur relative de l'énergie dans certains canaux en vue d'effacer les pôles superflus, 2) la largeur de bande des maximums pour restituer les formants non détectés.

1) Dans le cas où Max.2 est compris entre 800 et 1300 Hz, on teste le rapport d'énergie Max.3/Max.2. Si celui-ci est plus grand qu'un certain seuil, il résulte que l'énergie de Max.2 n'est pas suffisante pour constituer un pôle de la fonction de transfert du conduit vocal (fig. 1).

On relève empiriquement que le rapport d'énergie Max.3/Max.2 est en général supérieur à 1.2 ; s'il est compris entre 1 et 1.2, on doit tester l'émergence de Max.2 au-dessus du premier minimum d'énergie à droite de Max.2. Dans le cas où ce rapport est inférieur à 1,5Max.2 ne constitue jamais un pôle formantique.

Exemple d'application : Ech. 36, voyelle [i] de sympathique

Canaux	:	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Energie	:	13	9	4	5	4	4	6	9	8	6	8	9	12	9
Maximums			!		!				!					!	
détectés	:		400		957				2004					3420	

E (Max.3/Max.2) = 9/5 = 1.8

Maximums après correction : Max.1 400 Max.2 2004 Max.3 3420

Cet algorithme très simple aboutit au résultat désiré dans tous les cas où le système détecte des maximums non prévus par la fonction de transfert.

2) On aura remarqué que Max.3 et Max.4 deviennent Max.2 et Max.3, mais qu'ils ne sont pas affectés à des formants particuliers. La correspondance entre maximums et formants est établie à la sortie du deuxième algorithme (ALGO II).

A l'entrée d'ALGO II (fig. 2), on cherche à identifier les voyelles sur lesquelles F1 et/ou F2 ont été mal détectés : les voyelles sur lesquelles

F1 et F2 sont très voisins et confondus par le système, par exemple [a], [ɔ] etc..., celles sur lesquelles F2 et F3 forment un seul maximum, par exemple [à], [ɛ] etc... Puisque le maximum correspondant à F2 n'est pas détecté, c'est le maximum de F3 qui généralement prend la place de Max.2.

Ce sont les voyelles non fermées qui présentent ces anomalies. Par conséquent, les voyelles qui ont un Max.1 $\geq$ 450 (ouvertes) et un Max.2 $\geq$1900 (1) sont traitées dans une partie de l'algorithme qui doit permettre d'extraire F2 de Max.1 ou de Max.2.

Certains chercheurs (Shafer and Rabiner, 1970) ont proposé une technique pour résoudre le problème posé par la proximité des formants ; elle consiste en une analyse en bande étroite par la transformée en z de Chirp.

Mais on peut aboutir à un résultat au moins aussi satisfaisant, semble-t-il, à l'aide d'une procédure beaucoup plus simple.

a) On teste, dans ce but, la largeur de bande de Max.1 que l'on approxime en prenant en compte le nombre de canaux, entre deux minimums, dont on évalue le niveau relatif d'énergie. On s'aperçoit en effet que lorsque F1 et F2 se trouvent inclus dans Max.1, le nombre de canaux dans le Maximum est $\geq$ 4 et que le niveau dans les canaux est à peu près étalé.

Ex. : Echantillon 12, voyelle [a] dans rapide

Canaux	:	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Energie	:	12	14	13	12	11	11	8	6	7	7	5	3	4	7
Maximums				!						!				!	
détectés	:			1014						2569				3814	

Si dans 4 canaux au moins l'énergie est égale ou supérieure au seuil calculé, Max.1 contient F1 et F2 qui sont évalués à partir du centre de gravité des $\frac{n}{2}$ canaux de gauche (F1) et des $\frac{n}{2}$ canaux de droite (F2).

Dans l'exemple précédent, les formants calculés sont : F1 = 650 Hz et F2 = 1000 Hz. L'algorithme, dans sa première version, prévoyait ensuite l'affectation de Max. 2 à F3 et de Max.3 à F4 ; cette affectation automatique qui donne généralement de bons résultats, comme dans l'exemple précédent, entraîne toutefois un certain nombre d'erreurs, car il arrive que le pôle de F3 ne soit pas détecté : dans ce cas Max.2 représente F4.

Nous avons prévu de compléter cette partie de l'algorithme, dans une version ultérieure, afin d'éviter cette fausse affectation.

b) Soit maintenant l'exemple suivant :

Ex. : Echantillon 36, voyelle [a], dans délicate

Canaux	:	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Energie	:	12	12	7	6	6	7	10	9	9	6	4	7	10	7
Maximums				!					!					!	
détectés	:			735					2110					3563	

Le test sur la largeur de bande de F1 est négatif. On calcule alors la largeur de bande du Maximum suivant, c'est-à-dire Max.2, et on utilise dans ce

(1) Il y a de fortes chances pour que F2 soit mal détecté lorsqu'à un F1 de voyelle ouverte est associé un F2 supérieur à 1900.

but un coefficient différent de celui qui est utilisé pour l'évaluation de la largeur de Max.1. En effet, la répartition de l'énergie dans les canaux n'est pas la même pour Max.1 et Max.2 : en particulier elle est plus régulière dans Max.2, ce qui suppose l'utilisation d'un critère plus exigeant. Si le coefficient dépasse un certain seuil, défini empiriquement à partir des résultats de l'analyse, F2 et F3 sont représentés respectivement par la fréquence du centre de gravité des $\frac{n}{2}$ canaux de gauche et des $\frac{n}{2}$ canaux de droite de Max.2, et Max.3 = F4.

Ex. Voyelle [a] dans <u>délicate</u>

Coefficient de Max.2 = 14.28 5 (seuil)
F2 = 1700 Hz
F3 = 2300 Hz

Il est intéressant de constater que cette méthode permet de restituer un deuxième formant qui correspond à la valeur de la transition au contact de la consonne [t].

Dans le cas où le coefficient de Max.2 est inférieur au seuil, Max.2 = F2; si Max.3 < 3000, Max.3 = F3, dans le cas contraire F3 est égal à la fréquence des $\frac{n}{2}$ canaux de gauche du dernier Max. La restitution de F3 pour cette dernière opération ne donne pas toujours les résultats escomptés. Nous avons prévu la modification de cette partie de façon à identifier les antirésonances éventuelles dans la région de F3.

3) Dans la première partie de l'algorithme nous avons traité les voyelles non fermées qui étaient censées présenter une anomalie au niveau de F2. Dans une deuxième partie nous prenons en compte en même temps les voyelles ouvertes dont F1 et F2 semblent être représentées par Max.1 et Max.2 (Max.1 > 450 et Max.2< 1800), les voyelles fermées antérieures (Max.1 < 450 et Max.2 > 1800), et les voyelles fermées postérieures (Max.1 < 450 et Max.2 < 1800. Dans cet ensemble on retrouvera également les voyelles (i,y) sur lesquelles Max.2 aura été effacé par ALGOI. Pour ces dernières,on calcule le coefficient de largeur de bande de Max.2 (ex. Max.3), car le plus grand nombre d'erreurs provient, dans leur cas, de la mauvaise détection de F2 ou de F3.

Ex. Echantillon 21, voyelle [i], dans <u>solitude</u>.

Canaux	:	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Energie	:	15	10	6	7	6	6	9	11	8	8	6	8	12	8
Maximums			!		!				!					!	
détectés	:	491			957				2069					3567	
ALGO I	:				—										
ALGO II	:							Coef. de Max.2 (2069) = 6 > 5 (seuil)							
			!						!		!			!	
Formants	:	491							1900		2400			3567	

Ensuite on suppose provisoirement que Max.1 et Max.2 des autres voyelles de cet ensemble (ouvertes, fermées, antérieures et postérieures) représentent respectivement F1 et F2. On teste alors Max.3 : si Max.3 > 2000, l'hypothèse est vérifiée et Max.1 = F1, Max.2 = F2 et Max.3 = F3. Dans le cas contraire, Max.3 ne peut pas correspondre à F3 et comme on a par ailleurs un Max.2 non effacé, F2 se trouve à l'intersection de ces deux maximums. On rencontre ce type de détection sur les voyelles postérieures (o, u) caractérisées par un F2 bimodal.

Ex. Echantillon 12, voyelle [ɔ], dans solitude

Canaux	:	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Energie	:	13	13	9	11	8	10	6	3	4	6	3	5	7	3
Maximums			!		!		!				!			!	
détectés	:		527		950	!	1501				2518			3530	
Formants	:		527			1225					2518			3530	

Cette dernière partie de l'algorithme qui donne des résultats très satisfaisants ne conduit pas toujours à une affectation correcte de F3.

CONCLUSION

Le modèle, dans sa version actuelle, offre une méthode simple de détection des formants à partir des indices relatifs à la forme de la fonction de transfert, à la largeur de bande des maximums et au rapport entre les formants. Il permet de corriger 86 % des erreurs. Etant donné que ces erreurs s'élèvent à 50 % dans l'analyse spectrale par 14 canaux, l'erreur résiduelle, après le passage de l'algorithme, se limite à 7 %. Nous pensons, grâce aux modifications apportées, réduire considérablement ce taux d'erreurs, imputable essentiellement, comme on l'a vu, à une affectation incorrecte de F3, mais aussi aux valeurs-seuils de classification à l'entrée d'ALGO II. Ces dernières valeurs sont remplacées par un algorithme qui permet d'appliquer le même traitement à la classe des voyelles ouvertes anormales et seulement à cette classe.

REFERENCES

ATAL, B.S., HANAUER, S.L., Speech analysis and synthesis by linear prediction of the speech wave, J.A.S.A., 50, 1971, pp. 637-655.

BELL, C.G., FUJISAKI, H., HEINZ, J.M., STEVENS, K.N., HOUSE, A.S., Reduction of speech spectral by analysis -by- synthesis techniques, J.A.S.A., 33, 1961, pp. 1725-1736.

COKER, C.H., Real-time formant vocoder, using a filter bank, a general purpose digital computer and an analog synthesizer, J.A.S.A., 38, 1965, p. 940 (A).

FANT, G., Acoustic theory of speech production, La Haye, Mouton, 1960.

ITAKURA F., SAITO, S., An analysis-synthesis telephony based on maximum likelihood method, Proc. VI Int. Congr. Acoustics, C.5-5, C.17-20, Tokyo, 1968.

OLIVE, J.P., Automatic formant tracking by a Newton-Raphson technique, J.A.S.A., 50, 1971, pp. 661-670.

RABINER, L.R., SCHAFER, R.W., RADER, C.M., The chirp z-transform algorithm and its application, Bell System Techn. Journal, 48, 1969, pp. 1249-1292.

SCHAFER, R.W. and RABINER, L.R., System for automatic formant analysis of voiced speech, J.A.S.A., 47, 2, 1970, pp. 634-648.

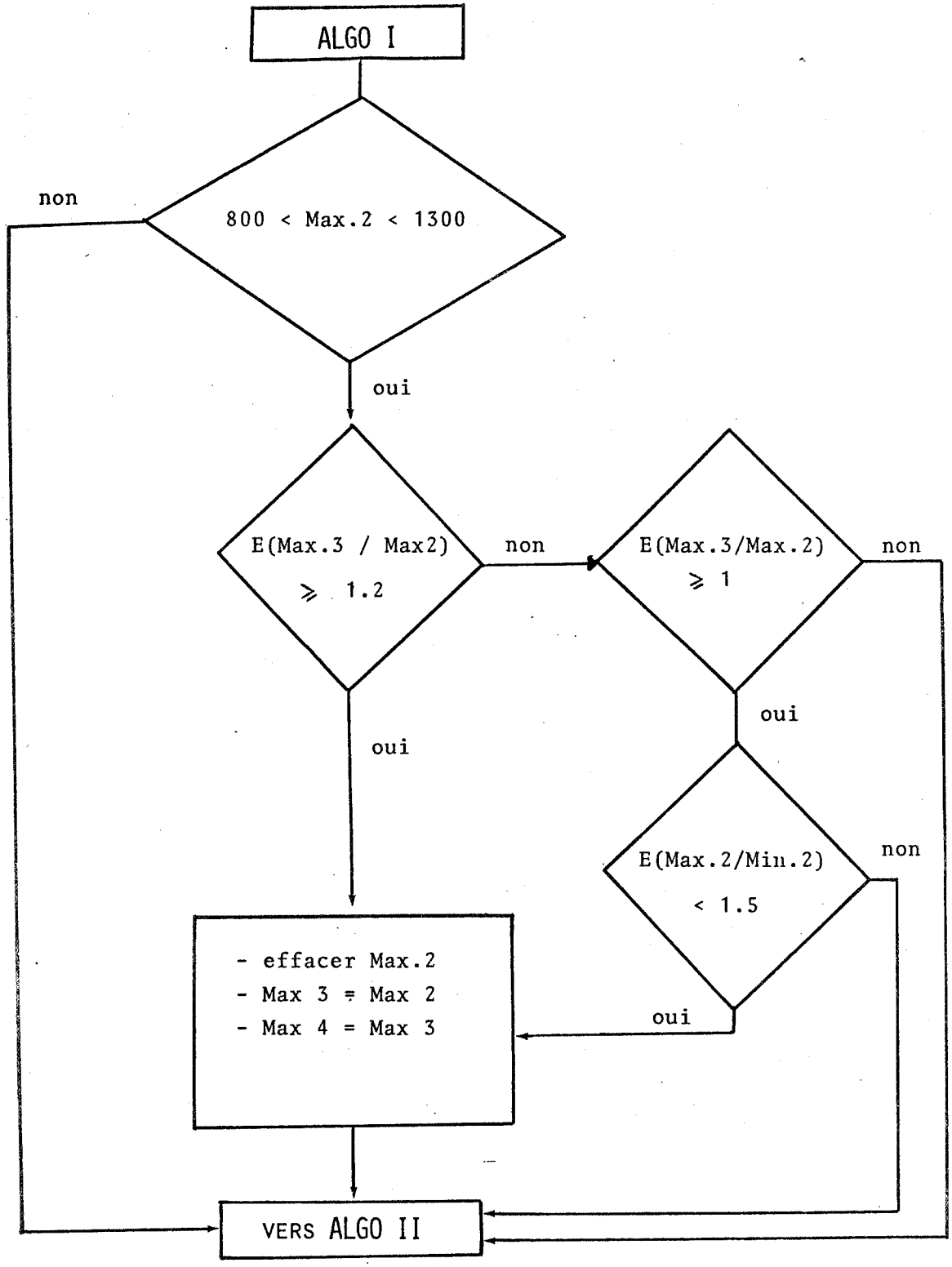


FIGURE 1

ALGORITHME D'EFFACEMENT DES PÔLES SUPERFLUS.

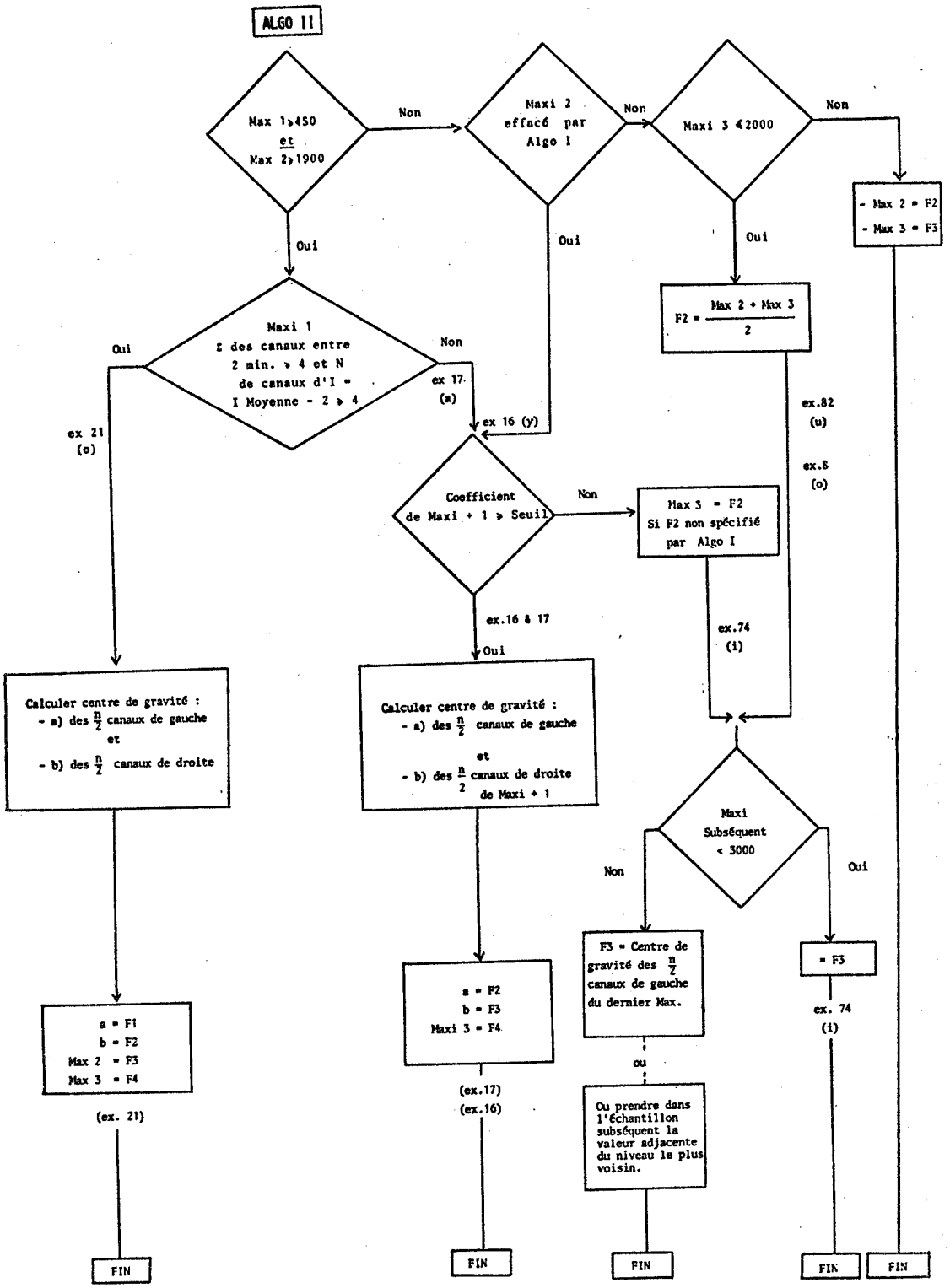


FIGURE 2

ALGORITHME DE RESTITUTION DES FORMANTS ET DE CADRAGE DES MAXIMUMS

PHRASE METHODE
MASCHVOIX = 49
FEMIVOIX = 0

NOMBRE D ECHANTILLONS 66 PITCH ENER PENT VARIAN CDG FB FN PLAN PHONEMES

CALCUL DES FORMANTS DU MOT : METHODE

NO SYLL	MAXIMUM-1	MAXIMUM-2	MAXIMUM-3	MAXIMUM-4	MAXIMUM-5	DEB VOY	CENTRE	FIN VI
1	613,41	1796,67	3569,23	,00	,00	5		
1	613,41	1796,67	3569,23	,00	,00	5		
Maximums	793,10	1891,18	2620,00	3580,36	,00		6	
Formants	793,10	1891,18	2620,00	3580,36	,00		6	
	607,14	1178,57	1878,57	3271,25	,00		7	
	607,14	1878,57	2667,65	,00	,00		7	
	760,27	1894,83	2833,93	3716,00	,00		8	
	760,27	1894,83	2833,93	3716,00	,00		8	
	673,33	1794,59	2776,79	3722,22	,00			9
	673,33	1794,59	2776,79	3722,22	,00			9
2	522,97	957,76	1583,12	2525,00	3537,50	22		
2	522,97	1583,12	2525,00	3537,50	,00	22		
	535,00	951,52	1707,00	2670,69	3524,44		23	
	535,00	1707,00	2670,69	3524,44	,00		23	
	719,05	1571,79	2521,43	3520,24	,00		24	
	719,05	1571,79	2521,43	3520,24	,00		24	
	544,87	945,71	1979,17	3530,26	,00		25	
Erreur d'affectation de F3	544,87	1462,44	3163,89	,00	,00		25	
	544,74	1265,83	2489,29	3530,56	,00		26	
	544,74	1265,83	2489,29	3530,56	,00		26	
	539,19	1339,34	2650,00	3515,52	,00		27	
	539,19	1339,34	2650,00	3515,52	,00		27	
	539,19	1322,50	2670,00	3518,00	,00		28	
	539,19	1322,50	2670,00	3518,00	,00		28	
Erreur d'affectation de F3	533,33	1338,79	3275,61	,00	,00		29	
	533,33	1338,79	3275,61	,00	,00		29	
	526,47	1337,07	2670,00	3537,50	,00		30	
	526,47	1337,07	2670,00	3537,50	,00		30	
	526,47	1234,18	2545,00	3543,10	,00			31
	526,47	1234,18	2545,00	3543,10	,00			31

NOSYL NOECH ENERGIE PAR CANAL RAPPORT ENERGIE/FREQUENCE
1 5 12 12 9 8 8 9 11 7 7 6 6 7 9 6

FIGURE 3

9èmes JOURNEES D'ETUDE SUR LA PAROLE

LANNION 31 mai - 2 juin 1978

ETUDE DES VARIATIONS ACOUSTIQUES DE LA VOYELLE DANS LES MONOSYLLABES EN FRANCAIS

Marie-Thérèse ROTH

Centre National d'Etudes des Télécommunications
Département Services Spéciaux du Téléphone
Lannion

.**RESUME**

Cette communication présente les tout premiers résultats d'une étude sur l'évolution dans le temps des zones de concentration d'énergie dans les monosyllabes en français.

Cette étude s'inscrit dans le cadre plus général des recherches menées au CNET sur la Reconnaissance Automatique de la Parole et nous pensons que cette première étape fournira une base de référence pour les études sur les voyelles dans des contextes plus larges tels que le mot et la phrase.

Pour cela, nous avons établi un corpus limité de 144 monosyllabes isolés de type C.V.C. qui a été lu par un locuteur et non pas recueilli à partir d'une conversation spontanée. Quatre voyelles et dix-sept consonnes ont été choisies pour la construction des monosyllabes. Les résultats obtenus sont ceux de la voyelle /a/ étudiée dans dix-huit syllabes.

Cette étude s'effectue selon deux points de vue : le point de vue paradigmatique qui donne une appréciation de la voyelle en elle-même (le même environnement consonantique apparaît immédiatement avant et après la voyelle : C_1VC_1), le point de vue syntagmatique qui nous guide dans l'étude des influences contextuelles..

Pour le moment, il s'agit de déterminer si la grandeur des variations constatées sur la voyelle /a/ à la fois à partir du vocodeur à canaux et du sonographe (comparaison des matériels) est due au contexte.

. SUMMARY

This paper presents the very-first results of a study on the steady-states and transitions of formant frequencies of the vowel /a/ in monosyllables.

This research aims to provide a reference basis for the study of vowels in larger contexts such as words and full sentences ; this should lead to the development of phonological rules that will be incorporated in the KEAL speech understanding system. Answers are sought for the following questions :

- is the influence of surrounding consonants upon the steady-state of the vowel significant and systematic enough to be taken into account in a Recognition Program, at least for French ?

- do the direction of the formant transitions from the vowel toward the consonant carry sufficiently discriminating information to be used as an aid for the automatic identification of the consonant ?

A list of 144 CVC type syllables (see annex I) were read by a native French speaker. The syllables were analysed by a sonograph (see Fig.I and annex 2) and 14 channel vocoder (see Fig.2 and annex 2). Figure 8 presents the results obtained by visual examination of sonogram. Figure 2 depicts the segmentation method for selecting the three representative vocoder samples for the vowel (beginning, steady-state and end). Figure 3 illustrates the selection of the energy peaks for a vocoder sample and Figure 7 shows the results of analysis of syllables from the vocoder data.(compare Figures 7 and 8).

The conclusions of this limited study are discussed in this paper. Vowel steady-state definition is constant into stop and fricative contexts. But, in a Recognition System, it is necessary to give another vowel steady-state definition for /a/ in nasal and /r/ context (the first peak disappears in /mam/ and /rar/). The direction of formant transitions sustains out the influence of the immediately preceding consonant upon the following vowel and the influence of the following consonant upon the immediately preceding vowel. The single second formant transition criterium does not allow to distinguish between the three places of production : labial, dental and velar. It allows only to eleminate some consonant and vowel arrangements.

Further research, including more speakers and larger linguistic context is in progress.

P L A N

INTRODUCTION

I - PROCEDURE EXPERIMENTALE

A) Corpus, locuteur, conditions d'enregistrement, matériels utilisés

B) Interprétation des données du sonographe et des vocodeurs à canaux

1) Les données du sonographe
2) Les données du vocodeur

a) Format des données
b) Sélection des échantillons représentatifs des débuts, fin et zones stables des voyelles : la segmentation
c) Sélection des canaux correspondants à des maxima d'énergie dans les échantillons choisis

II - PREMIERS RESULTATS ET DISCUSSION

A) La zone stable de /a/

1) Remarques sur l'influence de l'environnement consonantique sur la voyelle /a/
2) La voyelle /a/ dans les contextes occlusif et fricatif
3) La voyelle /a/ dans les autres contextes consonantiques

a) Le contexte nasal
b) Le contexte liquide

B) Etude des transitions

1) Remarques
2) La théorie

III - CONCLUSION

IV - BIBLIOGRAPHIE

V - ANNEXES 1 ET 2

VI - FIGURES

INTRODUCTION

Cette communication présente nos tout premiers résultats d'une étude sur l'évolution dans le temps des zones de concentration d'énergie dans les monosyllabes en français. Une seule voyelle, la voyelle /a/ a été analysée dans dix-huit monosyllabes isolés de type C.V.C. qui ont été lus par un locuteur masculin.

Cette étude s'inscrit dans un cadre plus général d'une recherche qui se poursuit actuellement sur un corpus plus large avec plusieurs locuteurs.

Pour le moment, nous allons essayer de déterminer si la grandeur des variations constatées sur la voyelle /a/ à la fois à partir du vocodeur à canaux et du sonographe (comparaison des matériels) est due au contexte.

Pour cela, nous exposerons tout d'abord notre procédure expérimentale, puis nos premiers résultats.

I - PROCEDURE EXPERIMENTALE

A) Corpus, locuteur, conditions d'enregistrement, matériels utilisés.

Notre but étant d'étudier la voyelle, nous avons limité notre corpus à 144 monosyllabes isolés, de type C.V.C. uniquement. Il permet d'analyser la voyelle en elle-même quand le même environnement consonantique précède et suit immédiatement la voyelle ainsi que l'influence consonantique sur la voyelle (phénomènes de coarticulation). A l'inverse, le type V.C.V. propose d'étudier de la consonne elle-même et l'influence vocalique sur la consonne.

Pour construire notre corpus, nous avons choisi quatre voyelles /a,i,u,ã/ et dix-sept consonnes /p,t,k,b,d,g,f,s,ʃ,v,z,ʒ,m,n,ɲ,l,r/. Trois des quatre voyelles /a,i,u/ qui se trouvent chacune à l'une des extrémités du triangle vocalique ont été retenues pour éviter les problèmes de chevauchement des zones de formants. Les voyelles nasales dont quatre /ɑ̃,ɔ̃,ɛ̃ et œ̃/ font partie du système vocalique français n'ont guère été étudiées jusqu'à présent. C'est pourquoi, nous avons inclus dans le corpus, une voyelle nasale /ã/. (Voir la définition articulatoire de ces voyelles et de ces consonnes à l'annexe 1).

La combinaison de ces consonnes et de ces voyelles entre elles, a abouti à une liste de mots et de syllabes n'ayant aucun sens dans la plupart des cas, cela élimine au maximum l'effet du sens sur la prononciation des syllabes. De plus, nous avons représenté sous une forme orthographique particulière pour éliminer le plus possible l'influence de l'orthographe sur la prononciation des syllabes (HOUSE, A., et FAIRBANKS, G., 1953). Le problème de la prosodie n'est pas pris en compte ici. Les syllabes seront toujours en position accentuée.

Un locuteur masculin a lu la liste de syllabes qui lui ont été présentées sur des fiches cartonnées. Pour éviter les phénomènes d'assimilation et les influences de rythme en particulier qui auraient pu se produire à cause d'une grande ressemblance phonétique entre les syllabes, un intervalle de temps de trois à cinq secondes a été intercalé entre la prononciation de chacune des syllabes. Une courte période d'entraînement a eu lieu avant l'enregistrement par une lecture à haute voix d'une vingtaine de syllabes. Au cours de l'enregistrement, des pauses ont été ménagées, et les six premières syllabes émises après chaque pause n'ont pas été prises en compte dans l'étude. Quand une syllabe prononcée nous a paru non satisfaisante, nous l'avons présentée une deuxième fois après la lecture complète du corpus.

L'enregistrement a eu lieu à la chambre sourde du Département "Etudes et Techniques d'Acoustique(E.T.A.)"au C.N.E.T.

L'enregistrement obtenu a été analysé sur un vocodeur à 14 canaux ainsi que sur le sonographe 6061B de le Kay Electric Company.(Voir en annexe 2 les principales caractéristiques de ces deux matériels utilisés).

B) Interprétation des données du sonographe et du vocodeur à canaux.

1) Les données du sonographe.

Un sonogramme (figure 1) représente les images des résonances acoustiques produites par les organes de la parole dans trois dimensions : temps, fréquence et intensité. Le temps se lit de gauche à droite, la fréquence de bas en haut. Le degré de noirceur des ombres sur le papier donne une idée de l'énergie.

On voit les formants (zones de concentration maximale d'intensité) qui sont représentés par des bandes sombres, plus ou moins horizontales sur l'échelle linéaire des fréquences qui a été calibrée sur la figure tous les 500 Hz (le calibrage est donné à l'extrême gauche de la figure par des bandes sombres et claires horizontales alternées)(DELATTRE, P., 1951).

2) Les données du vocodeur à canaux.

a) Format des données.

La figure 2 représente les données relatives à une syllabe analysée par le vocodeur à canaux. Il s'agit de la syllabe [sat]. Chaque ligne du tableau se rapporte à un échantillon de parole (un toutes les 13,3 millisecondes). La colonne 1 donne le numéro de l'échantillon. Les colonnes 2 à 15 représentent les valeurs de l'énergie (codée de 0 à 15) dans chacun des quatorze canaux. La colonne 16 donne la fréquence calculée du fondamental (Fo). La colonne 19 représente l'énergie totale de l'échantillon, c'est-à-dire la somme des énergies des quatorze canaux.

La colonne 20 donne l'énergie lissée (somme de l'énergie sur trois échantillons).

La colonne 17 représente la "pente" d'énergie.

Quant à la colonne 18, elle donne la pente lissée (somme sur trois échantillons).

Sur la figure 2, nous avons indiqué à l'aide de flèches les échantillons retenus pour l'analyse (voir le paragraphe suivant).

→ indique les échantillons représentatifs sélectionnés des début et fin de la voyelle étudiée.

o→ indique l'échantillon représentatif sélectionné de la zone stable de la voyelle.

Les signes → et o→ correspondent aux maxima et au minimum de la courbe lissée de la pente.

b) Sélection des échantillons représentatifs des débuts, fins et zones stables des voyelles : la segmentation.

Pour notre étude, il était nécessaire de délimiter la voyelle dans le temps (c'est le problème de la segmentation) et d'en déterminer la partie stable (nucléus). Dans ce but, nous avons sélectionné pour chaque syllabe trois échantillons censés représenter le début, le nucléus et la fin de la voyelle.

Pour cette sélection, nous avons utilisé la valeur de la courbe lissée de la pente (MERCIER, G. 1975) donnée à la colonne 18 (voir figure 2). La courbe correspondante donne un indice du changement qui se produit entre deux échantillons spectraux consécutifs.

Bien que les spectres évoluent de façon continue d'un son à un autre (d'une consonne à la voyelle suivante et vice-versa), on peut observer des pics et des creux dans la "pente". Les pics correspondent à des changements rapides dans l'évolution temporelle du spectre, ils marquent les limites entre consonnes et voyelles ; les creux (faibles pentes) correspondent aux zones de stabilité d'énergie et indiquent le "nucléus" de la voyelle, (mais ils se trouvent aussi dans certaines consonnes).

C'est pourquoi nous avons sélectionné comme échantillons représentatifs des début et fin de la voyelle, les échantillons correspondant aux maxima de cette pente et comme échantillon représentatif de la zone stable, celui qui correspond au minimum absolu de la pente. Quand deux minima de même valeur absolue apparaissent dans la voyelle, nous avons choisi le premier. Le lecteur peut observer une telle segmentation sur la figure 2.

c) Sélection des canaux correspondant à des maxima d'énergie dans les échantillons choisis

Sur les échantillons sélectionnés, nous avons relevé les canaux dans lesquels l'énergie présentait un maximum local. La figure 3 illustrant une telle sélection, représente la zone stable de /a/ dans /pap/. La figure 3a représente les données du vocodeur à canaux, c'est à dire les valeurs de l'énergie dans chaque canal pour le huitième échantillon de la syllabe. La figure 3b illustre la sélection des canaux selon les maxima locaux d'énergie.

II - PREMIERS RESULTATS ET DISCUSSION

A) La zone stable de la voyelle /a/.

1) Remarques sur l'influence de l'environnement consonantique sur la voyelle /a/.

La figure 4 indique les canaux dans lesquels un pic d'énergie a été relevé pour les échantillons représentatifs de la zone stable de la voyelle /a/ dans tous les contextes consonantiques /p,t,k,b,d,g,m,n,ɲ,f,s,ʃ,v,z,ʒ, l,r/. Nous pouvons noter qu'une grande stabilité du nombre de pics et de leur position dans l'échelle des fréquences et de leur énergie se dégage de cette figure. Les cas de [mam] , [ɲaɲ] et [rar] seront discutés plus loin dans le texte.

Nous avons séparé l'étude de la voyelle /a/ dans les contextes occlusif et fricatif d'une part (fig 4a et 4b), de celle des autres contextes, nasal et liquide, d'autre part (fig 4c et 4d).

2) La voyelle /a/ dans les contextes occlusif et fricatif.

Quatre pics et quatre seulement ont été régulièrement détectés pour les douze consonnes concernées dans les canaux indiqués dans le tableau 1.

Pics	Canaux	Fréquence approximative
1er pic	3	750
2e pic	4,5 - 6,5	1062 - 1600
3e pic	9 - 10	2350 - 2650
4e pic	12 - 13	3300 - 3700

Tableau 1 : Répartition en fréquence des quatre pics caractérisant la voyelle /a/

Nous pouvons remarquer une très grande stabilité de la définition de la zone stable, quelles que soient les consonnes environnantes.

3) La voyelle /a/ dans les autres contextes consonantiques.

La remarque précédente reste également valable dans les environnements de /l/ et de /n/ (voir fig.4c et 4d).

En ce qui concerne les 3 autres environnements consonantiques nous observons les faits suivants :

a) le contexte nasal:

Dans l'environnement de /m/, le premier pic de la voyelle /a/ n'est pas visible, mais nous pouvons voir une concentration d'énergie importante dans les 6 premiers canaux (1) (figure 5a) et un déplacement vers la droite du 2e pic par rapport au contexte /n/. Cette différence dans la zone stable de la voyelle est évidente sur les sonogrammes schématisés de la figure 6 qui représentent [mam], [nan] et [ŋaŋ], sur lesquels nous avons indiqué les valeurs des antiformants pour les deux consonnes /m/ et /n/ : environ 1000 Hz pour /m/ et 1700 Hz pour /n/. Pour ces deux consonnes, Fujimura (FUJIMURA, O., 1962) donne ce même ordre de grandeur. Il indique également un antiformant de plus de 3000 Hz pour la nasale vélaire /ŋ/ que le français ne connait qu'en position finale dans des mots ou des suffixes d'origine anglaise du type "parking", "living" etc..., mais nous n'avons pas encore à notre connaissance de données précises sur la valeur de l'antiformant de /ɲ/ sans doute plus élevée que celles de /m/ et de /n/.

(1) M. ROSSI travaille actuellement sur la restitution automatique de formants. Voir également l'article de M.LENNING (LENNING, M., 1977)

Ces antiformants sont sans doute responsables du relèvement du deuxième formant (F_2) dans [nan] et [ɲan] par rapport à [mam]. Nous pouvons également observer sur cette figure 6 la direction différente des transitions de formants dans les deux cas dont nous reparlerons un peu plus loin dans le texte.

b) le contexte liquide:

Nous pouvons observer la disparition du deuxième pic dans [ʒaʒ]. Ce cas est semblable à celui de /m/ que nous avons vu précédemment. Cette disparition que nous observons sur les données du vocodeur à canaux est due à un rapprochement important des deux premiers formants F_1 et F_2. Cette observation est illustrée par la figure 5b.

La difficulté de distinguer entre les deux premiers pics dans les cas de [mam] et de [ʒaʒ], pourtant facilement différenciés sur le sonogramme, est due au manque de précision dans les basses fréquences du vocodeur utilisé. Ceci peut être illustré en comparant les figures 5a, 5b, 6 et 8.

B) Etude des transitions

Nous allons étudier maintenant la comparaison entre les échantillons des début et fin de la voyelle /a/ et l'échantillon de la zone stable.

La figure 7 représente les variations de la voyelle /a/, variations calculées entre les trois échantillons caractéristiques de la voyelle dans les différents contextes consonantiques. Cette figure permet de voir l'évolution dans le temps de la position des pics. Les premiers et derniers points représentent les début et fin de la voyelle, les points intermédiaires caractérisent la zone stable de la voyelle.

La figure 8 représente la même information obtenue à partir des sonogrammes. Il n'y a pas d'opposition pour les 3 premiers pics entre les deux sources d'information dont la première, celle du vocodeur, pourrait être obtenue automatiquement sauf pour les cas de [mam] et de [ʒaʒ] qui ont été précédemment expliqués. Cependant le degré de variation n'est pas toujours équivalent. C'est ce que nous pouvons constater en comparant l'évolution du troisième pic de /a/ dans [gag] que nous trouvons dans les données du vocodeur à canaux (fig. 7b) et celles relevées sur le sonogramme (fig. 8b).

1) Remarques

Nous pouvons nous attendre à une influence importante de la consonne qui suit la voyelle puisqu'elle appartient à la même syllabe (MALMBERG, B., 1955).

Cette influence retenue par Malmberg apparaît nettement sur les figures 7 et 8. L'influence de la consonne précédente sur la voyelle qui la suit immédiatement apparait clairement aussi bien sur les données du vocodeur à canaux, où l'on voit une asymétrie dans les variations, que sur les sonogrammes où un déplacement vers la gauche de la zone stable de la voyelle est très net.

2) La théorie

La théorie permet de prévoir le sens des transitions dans la voyelle à partir de la connaissance du contexte consonantique puisque l'évolution des formants dans la voyelle reflète le point d'articulation de la consonne, Liberman, (LIBERMAN, A., DELATTRE, P., GERSTMAN, L., COOPER, F., 1956). Nous

pouvons illustrer ceci en reproduisant la figure 1 de l'article cité.

Comme le prédit la théorie, nous observons une transition légèrement montante du second formant dans le contexte de consonnes labiales /p/ et /m/, la transition étant plate dans le contexte /b/, légèrement descendante dans le contexte des consonnes dentales /t/ /d/ /n/ /l/ /ʃ/ et /z/ (la transition apparaît plate sur les données du vocodeur sauf pour /l/ où la transition est descendante), et nettement descendante dans le contexte des consonnes palatales et vélaires /ɲ/, /k/ et /g/. Les transitions du second formant pour une durée assez courte (ou une vitesse rapide) sont sans doute les plus puissants indices pour la distinction entre les points d'articulation comme le remarque Delattre. (DELATTRE, P., 1958).

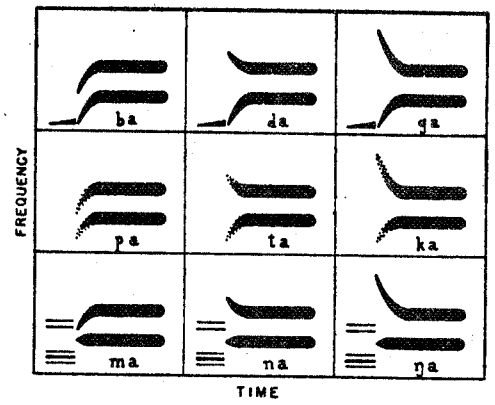


FIG. 1. Hand-painted spectrograms of nine consonant-vowel syllables, showing some of the principal acoustic cues for the perception of the stop and nasal consonants.

Cependant cette différence n'apparaît pas toujours nettement, surtout sur les données du vocodeur à canaux. Dans le contexte chuintant /ʃ/ et /ʒ/ le second formant est plat. Par contre sur le sonogramme, le deuxième formant descend légèrement, d'une façon plus claire dans le contexte /ʃ/ que dans le contexte /ʒ/. Pour les dentales sifflantes /s/ et /z/, on observe une légère montée plus une légère descente vers la zone stable de la voyelle sur le sonogramme (voir figures 7 et 8).

L'évolution de F_1 n'apportant aucune information, la direction des formants permettrait seulement d'interdire certaines combinaisons (ROSSI, M., 1975).

III - CONCLUSION

Nos conclusions concernent à la fois le matériel utilisé et les premiers résultats obtenus.

Du point de vue des matériels utilisés pour une étude de ce genre, le sonographe donne une information visuelle plus directe que le vocodeur à canaux, notamment du suivi des formants (voir l'analyse des deux exemples [mam] et [ʒa ʒ], tout au moins en ce qui concerne la voyelle /a/. Quant au vocodeur à canaux, il présente l'avantage d'indiquer l'énergie de façon quantitative, alors que le sonographe ne permet de juger qu'approximativement des valeurs d'énergie. Les deux matériels sont donc nécessaires dans le cadre d'une étude comme la nôtre, les résultats obtenus à partir du vocodeur à canaux étant plus directement applicables et pouvant être plus facilement intégrés dans les programmes de reconnaissance. Mais la présentation des résultats dépend du matériel utilisé (nombre des canaux, largeur de bande, etc...).

Du point de vue des résultats, les conclusions partielles sont les suivantes :

La définition des zones stables de la voyelle /a/ reste constante dans les contextes occlusif et fricatif, et il ne semble donc pas nécessaire de prendre en

compte le contexte dans un système de reconnaissance. Il semble à première vue qu'il faille donner une autre définition de la zone stable de la voyelle /a/, en contexte nasal et pour /r/, dans un système de reconnaissance qui devra prendre en compte entre autre la disparition éventuelle d'un des quatre pics (le premier dans [mam] et dans [ʁa ʁ]).

L'étude des transitions a confirmé l'existence théorique de l'influence de la consonne précédente sur la voyelle qui la suit immédiatement (Delattre) et aussi de l'influence de la consonne qui suit immédiatement la voyelle sur celle-ci si la consonne fait partie de la même syllabe que la voyelle (Malmberg). Cependant, les résultats prédits n'apparaissent pas clairement sur nos données. Conformément aux travaux entrepris par M. Rossi sur les voyelles françaises, la direction des transitions du 2e formant peut permettre d'éliminer certaines combinaisons de consonnes et de voyelles ou de voyelles et de consonnes, mais ce critère seul ne permet pas de distinguer entre les trois points d'articulation principaux (labial, dental et vélaire).

L'étude se poursuit actuellement d'une part en analysant le corpus sur d'autres matériels tels que le vocodeur numérique (comprenant un ensemble de filtres numériques programmables) et Ciphon (programme simulant un vocodeur à formants à trente-deux canaux fabriqué par la Société Thomson); d'autre part, elle se continue sur un corpus plus large avec plusieurs locuteurs puisqu'elle vise à l'étude de phrases et seuls les résultats qui restent pertinents au niveau de la phrase seront retenus.

IV - BIBLIOGRAPHIE

DELATTRE, Pierre. : "The Physiological Interpretation of Sound Spectrograms" (1951) pp 225-235 dans "Studies in French and Comparative Phonetics", M.I.T. Press, (1966)

"Les indices Acoustiques de la Parole" (1958) pp 276-286 dans "Studies in French and Comparative Phonetics", M.I.T. Press, (1966)

FUJIMURA,Osamu : "Analysis of Nasal Consonants" (1962) pp 238-248 dans "Readings in Acoustic Phonetics", M.I.T. Press, (1969)

HOUSE, Arthur.S. et FAIRBANKS,Grant : "The influence of Consonant Environment upon the Secondary acoustical Characteristics of Vowels" J.A.S.A. 25 ; pp 105-113.

LENNING, Matthew : "Selecting Syllable Nuclei in French".Paper presented at the 94th meeting of the Acoustical Society of America, Miami Beach, Décembre 1977.

LIBERMAN,Alvin M., DELATTRE, Pierre C., GERSTMAN, Louis J., COOPER, Franklin S., "Tempo of Frequency Change as a Cue for Distinguishing Classes of Speech Sounds" (1956) pp 159-169 dans " "Readings in Acoustic Phonetics" M.I.T. Press, (1969)

MALMBERG, Bertil : "The Phonetic Basis for Syllabic Division (1955) pp 293-300 dans "Readings in Acoustic Phonetics", M.I.T. Press, (1969)

MERCIER, Guy : "Automatic Segmentation of Speech into Syllabic and Phonemic units " : Application to French Words and

Utterances " pp 359-382, Academic Press London, 1975.

OSHIKA, Beatrice; ZUE Victor; WEEKS, Rollin; NEU, Helene; and AURBACH, Joseph, : "The Role of Phonological Rules in Speech Understanding Research", ICEE Transactions on Acoustics, Speech and Signal Processing, Vol.ASSP-23, February 1975, pp 104-12.

ROSSI, Mario : "Les contraintes phonologiques dans un système de Reconnaissance de la Parole", pp 160-190, J.E.P., Toulouse (1975),

REMERCIEMENTS:

Nous remercions Jacqueline Vaissière pour son aide dans la définition du sujet et dans la rédaction de cet article.
Que le département ETA, (en particulier Raymond Descout et Françoise Emerard) soit remercié d'avoir mis à notre disposition son sonographe et de nous avoir appris à nous en servir.
Merci également à Christian Le Corre qui a réalisé le traitement vocodeur de notre corpus.

Cette étude s'inscrit dans le cadre de la préparation d'une thèse de troisième cycle, qui sera présentée à l'Université de Haute-Bretagne, l'essentiel du travail étant réalisé au Centre National d'Etudes des Télécommunications, à Lannion.

ANNEXE 1

1) Définition articulatoire des voyelles utilisées

Trois types de définition pour les voyelles

a) antérieure/postérieure : /i/ et /u/
b) ouverte/fermée : /a/ et /i/ ; /a/ et /u/
c) orale/nasale : /a/ et /ã/

2) Définition articulatoire pour les consonnes utilisées

Trois types de définition

a) sourdes/sonores :- p,t,k,f,s,ʃ
- b,d,g,v,z,ʒ,m,n,ɲ,l,ʀ

b) mode d'articulation :
- les occlusives : p,t,k,b,d,g
- les fricatives : f,s,ʃ,v,z,ʒ
- les nasales : m,n,ɲ,
- les liquides : l, ʀ

c) point d'articulation
- les bilabiales : p,b,m
- les labiodentales : f,v
- les dentales : t,d,n,l,s,z
- les prédorsales : ʃ, z
- les palatales : ɲ
- les vélaires : k,g, ʀ

ANNEXE 2

1) Caractéristiques principales du sonographe

- Il s'agit du sonographe 6061B de la Kay Electric Company
- Le tempo de parole est de 2,4 secondes
- Choix du filtrage large (300 Hz) permettant de détacher les formants
- Le sonogramme qui est une représentation visuelle du spectre dans le temps est obtenu à partir d'un enregistrement direct de quelques secondes de parole
- Un stylet enduit de noir de fumée vient s'appuyer contre un tambour recouvert d'un papier spécial tournant à vitesse constante. Le stylet peint le papier en montant progressivement le long du tambour

2) Caractéristiques du vocodeur à canaux

- Nombre de canaux : 14
- Intervalle entre deux échantillons : 13,3 millisecondes

Les fréquences des canaux sont données dans le tableau suivant :

N° du canal	Fréquence centrale (Hz)	Largeur (Hz)
1	350	200
2	550	200
3	750	200
4	950	200
5	1175	250
6	1450	300
7	1750	300
8	2050	300
9	2350	300
10	2650	300
11	2950	300
12	3300	400
13	3700	400
14	4100	400

Le vocodeur est directement relié à l'ordinateur CII/1070 sur lequel les résultats sont traités par un programme.

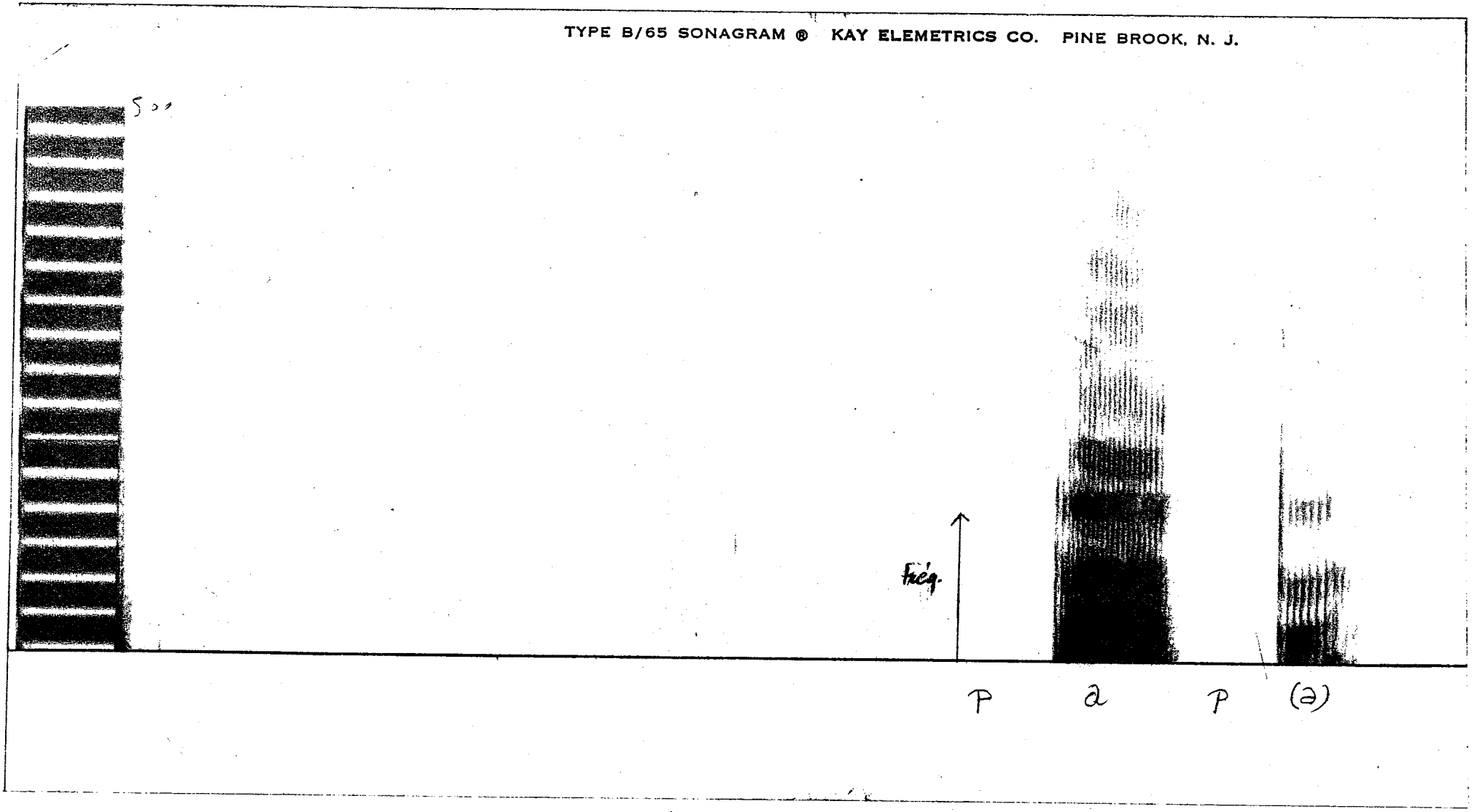


Figure I: *représentation du sonagramme de la syllabe* [pap].

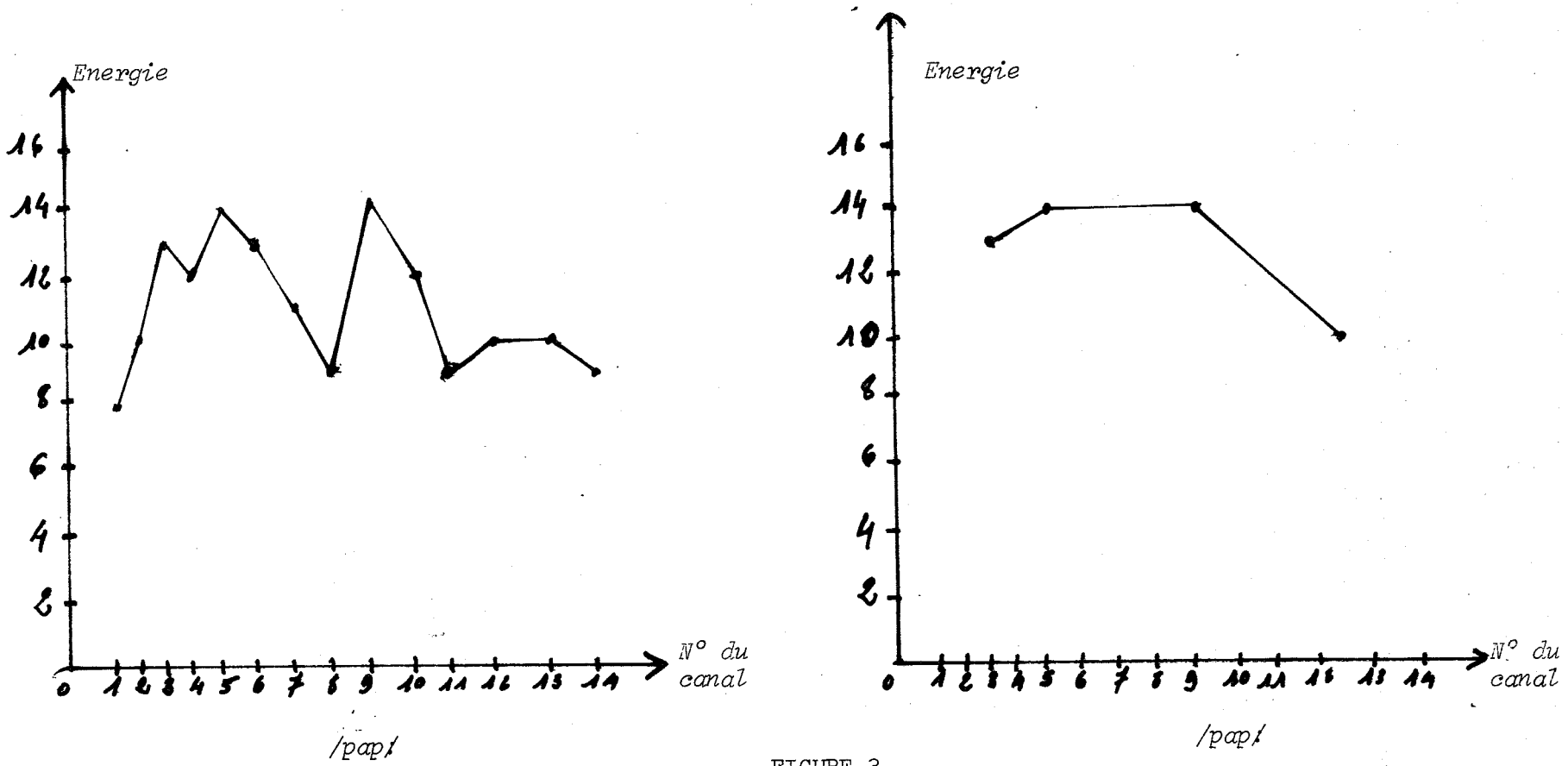


FIGURE 3

SELECTION DES SOMMETS D'ENERGIE DANS LES ECHANTILLONS REPRESENTATIFS DE LA VOYELLE

Fig.3a

Représentation de la répartition d'énergie dans les 14 canaux (exemple: zone stable de /a/)

Fig.3b

Sélection des canaux correspondant aux pics d'énergie (canaux 3,5,9 et 12,5 pour notre exemple)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
	53			SSAATT					1		1			26		1720				
									2		37			45		312				
1	0	0	0	0	0	0	1	0	2	2	1	1	1	2	0	7	0	10	0	
2	0	0	0	0	1	0	2	0	3	3	2	2	2	2	0	6	18	17	48	
3	0	0	0	0	0	0	2	3	4	4	2	2	2	2	0	5	16	21	62	
4	0	0	1	0	0	1	3	3	4	4	1	2	2	3	0	5	18	24	66	
5	0	0	0	0	0	1	3	3	4	3	2	1	2	2	0	8	21	21	74	
6	1	0	0	0	1	1	3	4	4	4	3	2	3	3	0	8	24	29	87	
7	1	0	0	1	2	2	3	4	6	5	4	3	3	3	0	8	25	37	107	
8	2	1	0	1	1	1	4	4	6	6	4	4	3	4	0	9	22	41	110	
9	0	0	0	1	0	1	3	3	5	6	3	3	3	4	0	5	21	32	108	
10	0	0	1	1	1	2	3	3	5	5	3	4	3	4	0	7	21	35	107	
11	1	1	0	1	1	2	4	5	5	5	4	4	3	4	0	9	29	40	110	
12	2	1	0	1	1	1	4	3	5	6	3	3	2	3	0	13	38	35	107	
13	2	2	2	0	1	1	3	1	4	5	2	2	3	4	0	16	82	32	109	
14	5	5	4	2	2	2	2	2	3	5	2	2	2	4	0	53	93	42	169	
début → 15	11	11	8	7	6	8	5	5	8	9	3	3	5	6	0	24	96	95	256	
16	11	13	11	8	8	10	6	6	10	11	6	5	7	7	114	19	49	119	348	
17	10	12	13	9	9	11	9	8	11	12	8	6	8	8	115	6	33	134	387	
18	10	11	13	9	11	11	10	8	11	12	7	6	8	7	116	8	22	134	406	
19	10	10	13	9	12	11	10	9	12	11	8	7	9	7	119	8	19	138	416	
milieu → 20	10	10	12	9	12	11	10	10	13	13	9	8	9	8	123	3	15	144	423	
21	10	10	12	9	12	11	10	9	13	13	9	7	8	8	124	4	18	141	426	
22	10	11	12	9	12	11	10	9	12	13	8	7	8	9	132	11	39	141	412	
fin → 23	10	11	12	8	10	10	9	8	11	12	8	6	7	8	137	24	119	130	377	
24	9	11	10	7	8	9	6	6	8	11	6	4	4	7	146	84	130	106	258	
25	6	8	0	2	0	4	0	0	0	1	0	0	1	0	146	22	114	22	128	
26	0	0	0	0	0	0	0	0	0	0	0	0	0	0	146	8	39	0	30	
27	0	0	2	0	0	0	0	0	2	3	0	0	0	1	0	9	18	8	9	
28	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	10	1	9	
29	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	
30	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
31	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
32	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
33	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
34	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
35	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
36	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	15	0	0	
37	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	15	69	0	15	
38	0	0	0	0	0	2	2	1	2	3	1	1	2	1	0	54	93	15	84	
39	3	3	5	2	4	6	6	4	7	8	5	5	5	6	0	24	102	69	139	
40	6	5	4	1	2	5	3	2	6	7	3	3	3	5	0	24	66	55	171	
41	9	8	3	2	1	5	1	1	5	6	0	1	4	1	189	18	58	47	163	
42	8	8	5	3	3	6	3	2	4	8	2	2	5	2	197	16	68	61	153	
43	5	6	5	2	3	5	2	2	1	7	2	0	3	2	192	34	58	45	117	
44	2	0	3	0	2	2	0	0	0	2	0	0	0	0	192	8	49	11	65	
45	3	1	0	0	1	2	1	0	0	1	0	0	0	0	0	7	0	9	0	
46	1	2	1	0	0	2	0	0	0	0	0	0	0	0	0	5	0	6	0	
47	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	
48	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
49	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
50	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
51	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
	Valeurs de l'énergie dans les 14 canaux															Fo	pente	E de l'énergie		

Figure 2: *données du vocodeur à canaux relatives à la syllabe [sat].*

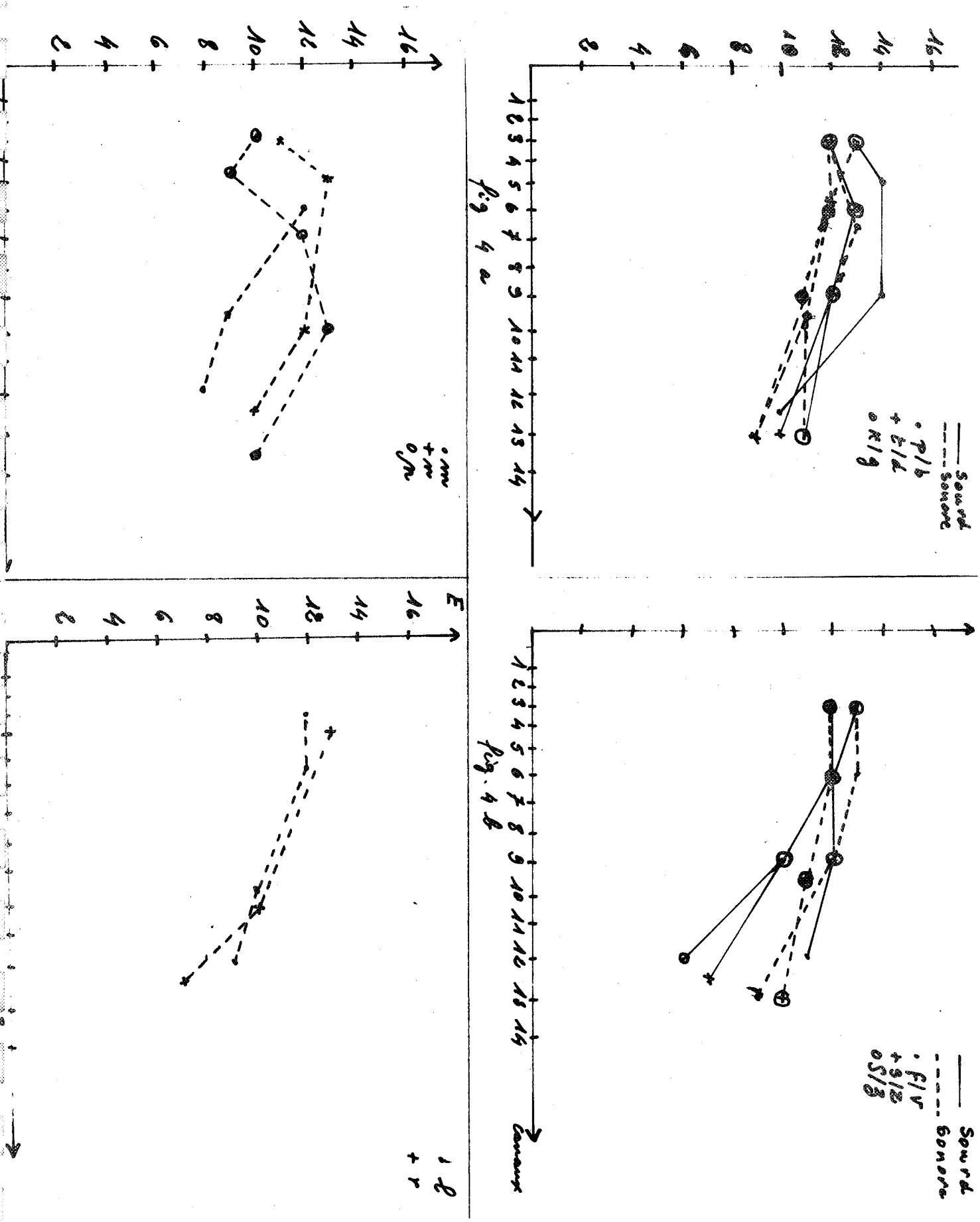


Figure 4: *influence des consonnes sur les pics d'énergie relevés dans la zone stable de la voyelle /a/.*

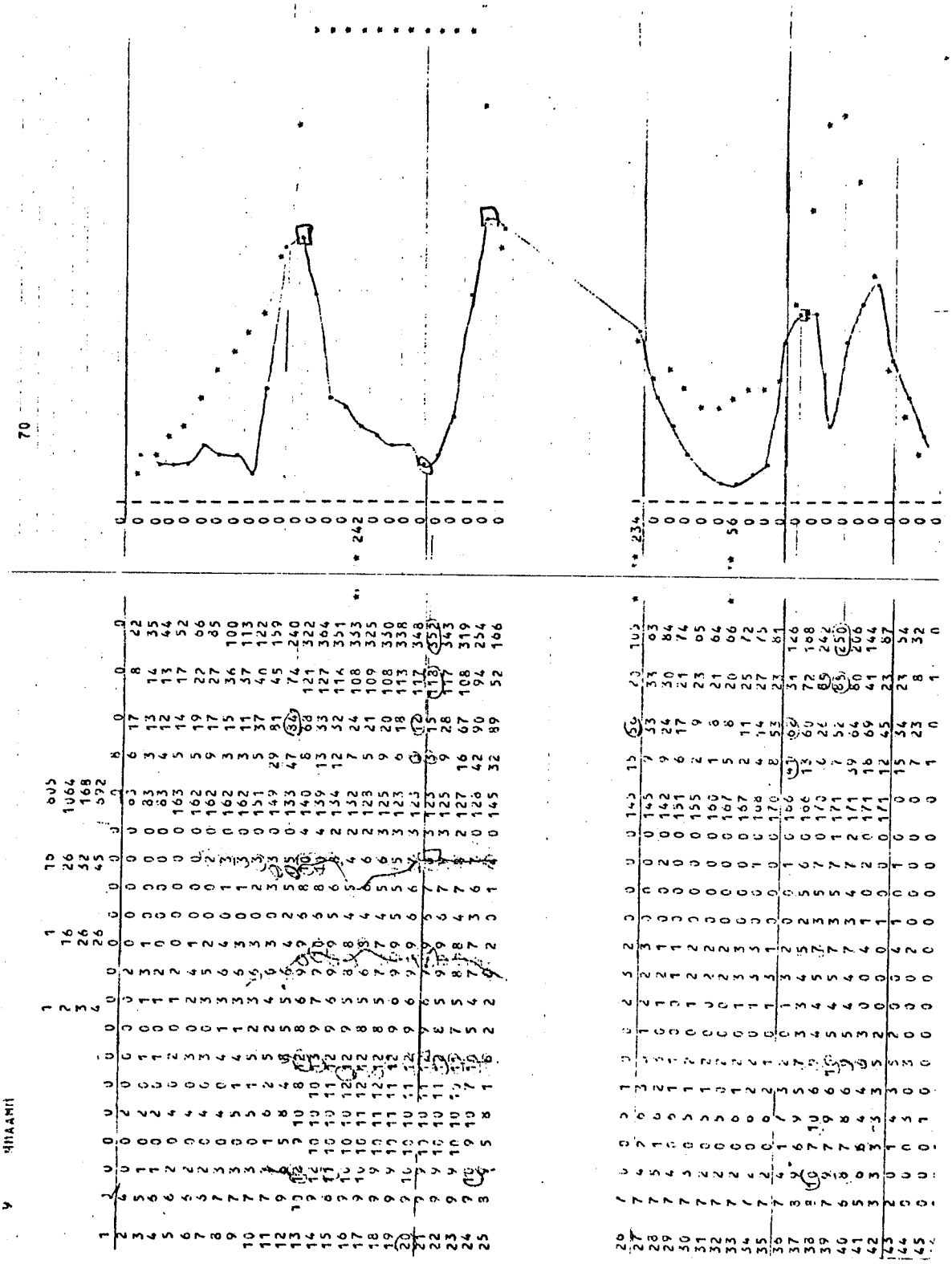


Figure 5a: montre une concentration d'énergie importante dans les six premiers canaux du vocodeur pour la voyelle /a/ dans l'environnement de /m/. Le premier pic n'est pas visible.

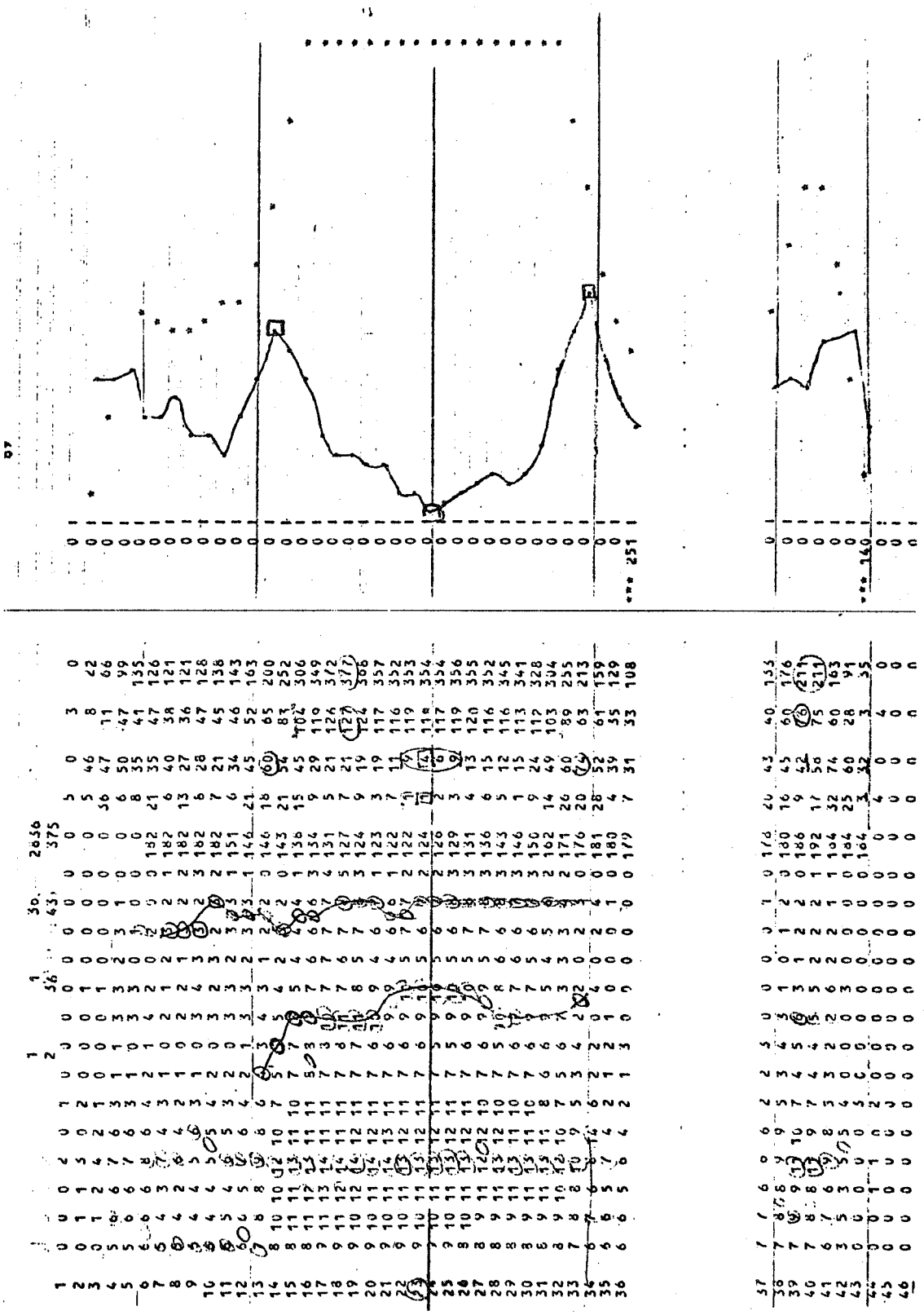


Figure 5b: montre la disparition du deuxième pic dans la voyelle /a/ et une concentration d'énergie dans les six premiers canaux du vocodeur.

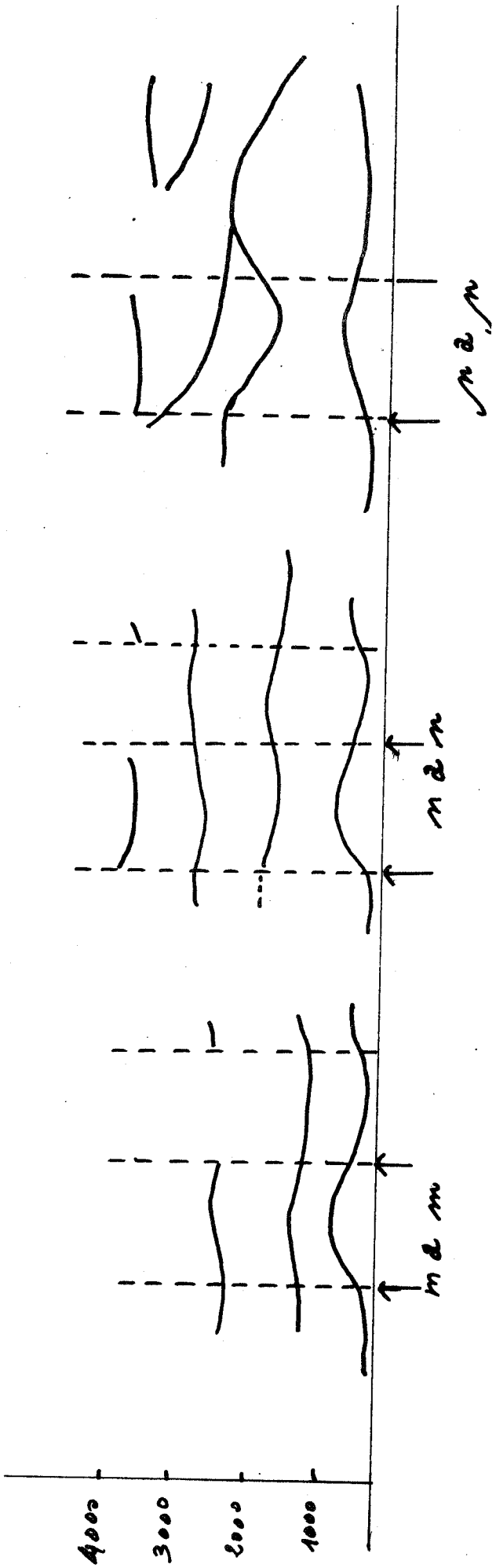


Figure 6: représente les données du sonagraphe pour la voyelle /a/ dans les contextes consonantiques /m/; /n/; /ɲ/.

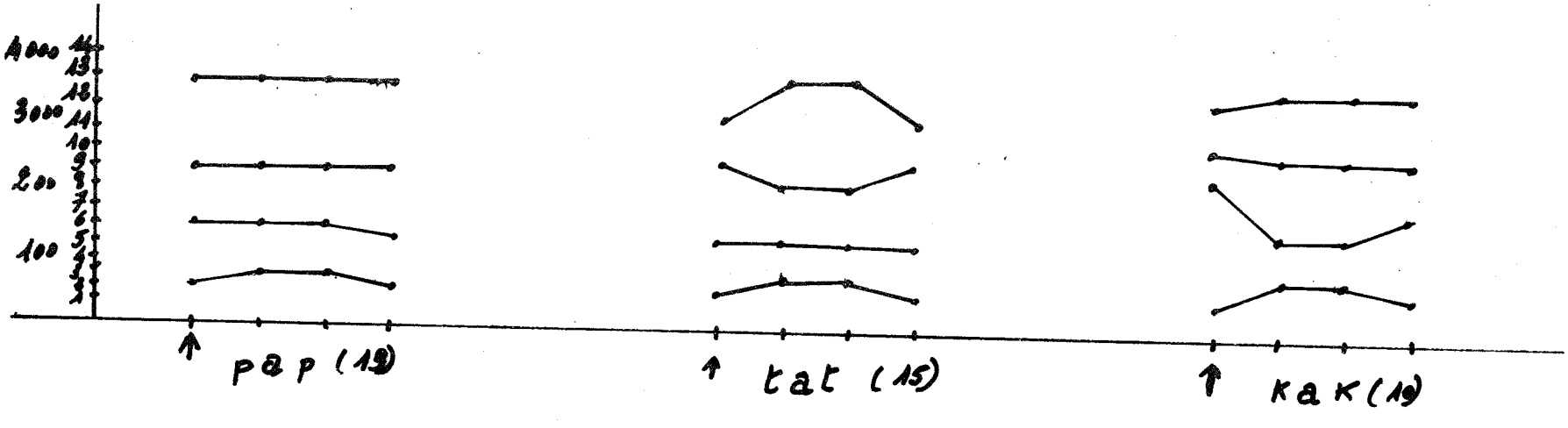


Fig.7a

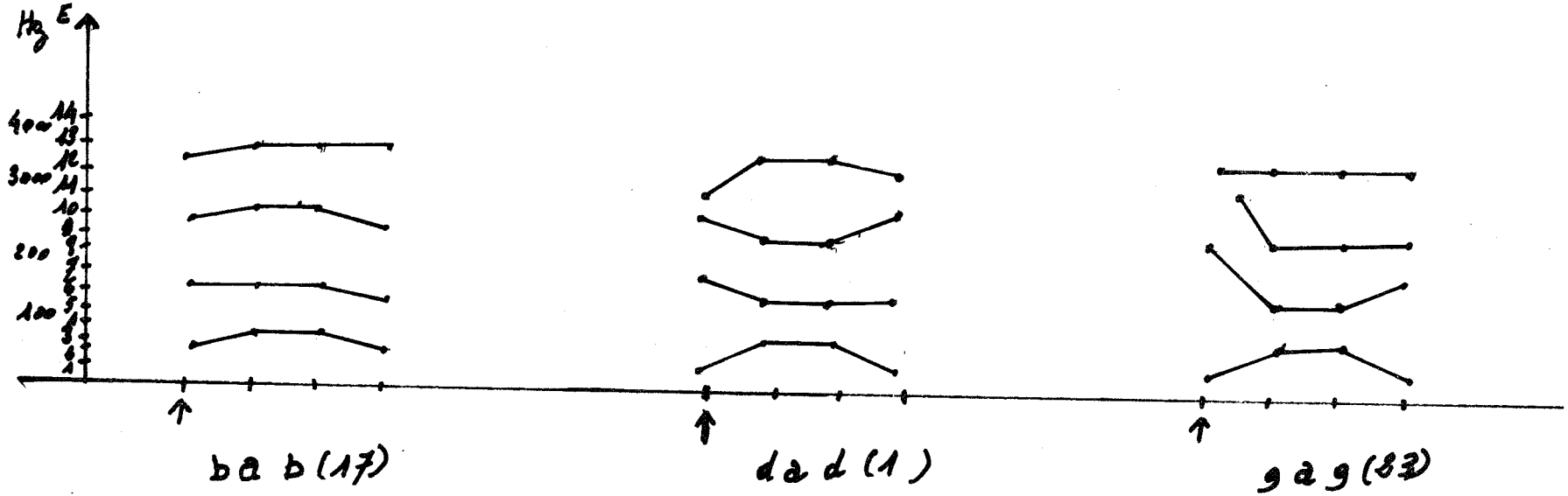


Fig.7b

Figure 7 montre l'évolution des échantillons caractéristiques de la voyelle/a/ en fonction du contexte consonantique. Les données sont celles du vocodeur à canaux.

Les figures 7a et 7b représentent la voyelle /a/ dans le contexte occlusif.

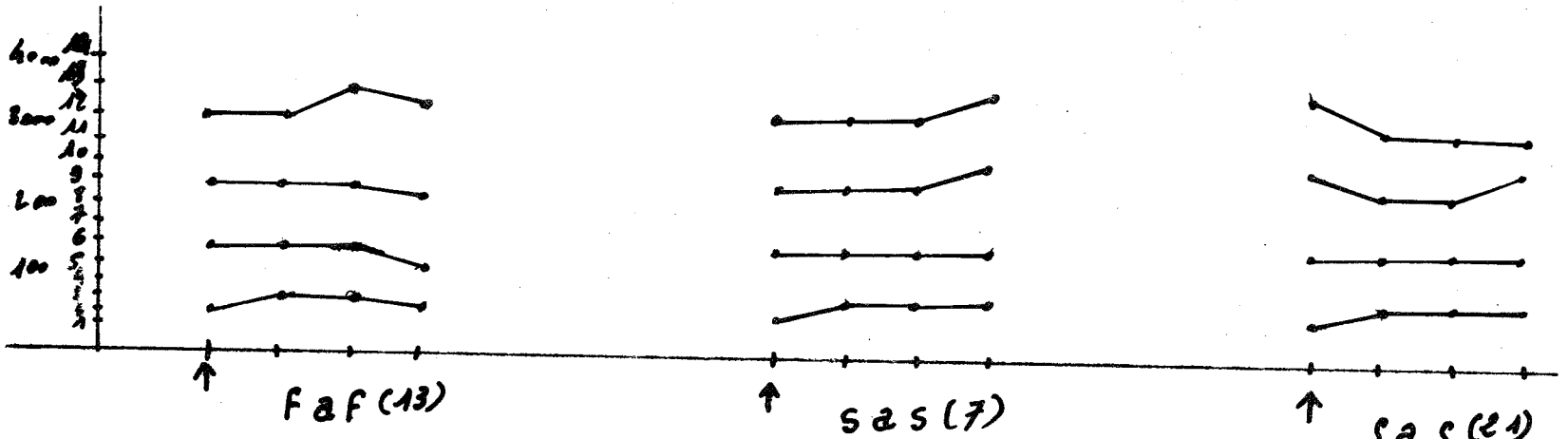


Fig. 7c

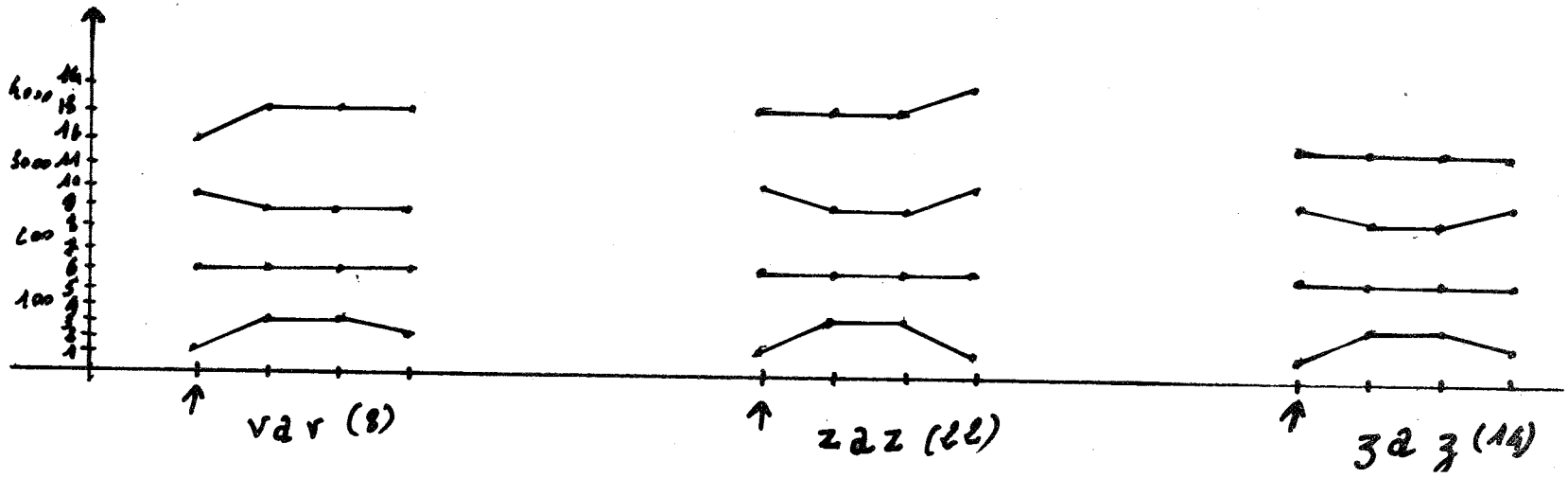


Fig. 7 d

Figures 7c et 7d: représentent la voyelle /a/ dans le contexte consonantique fricatif.

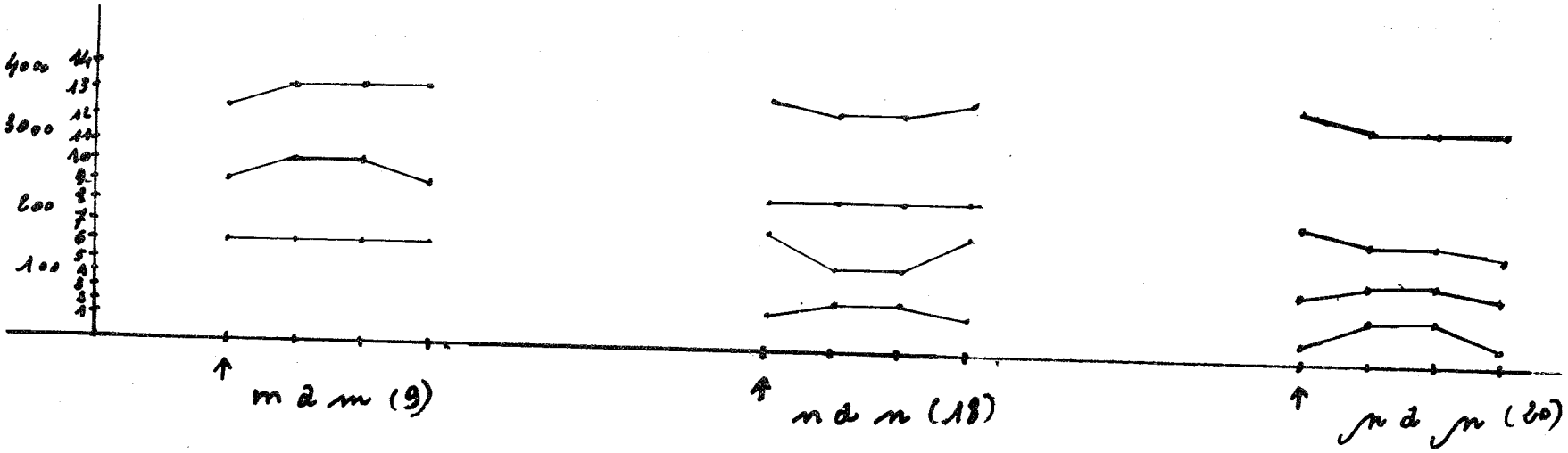


Fig. 7e

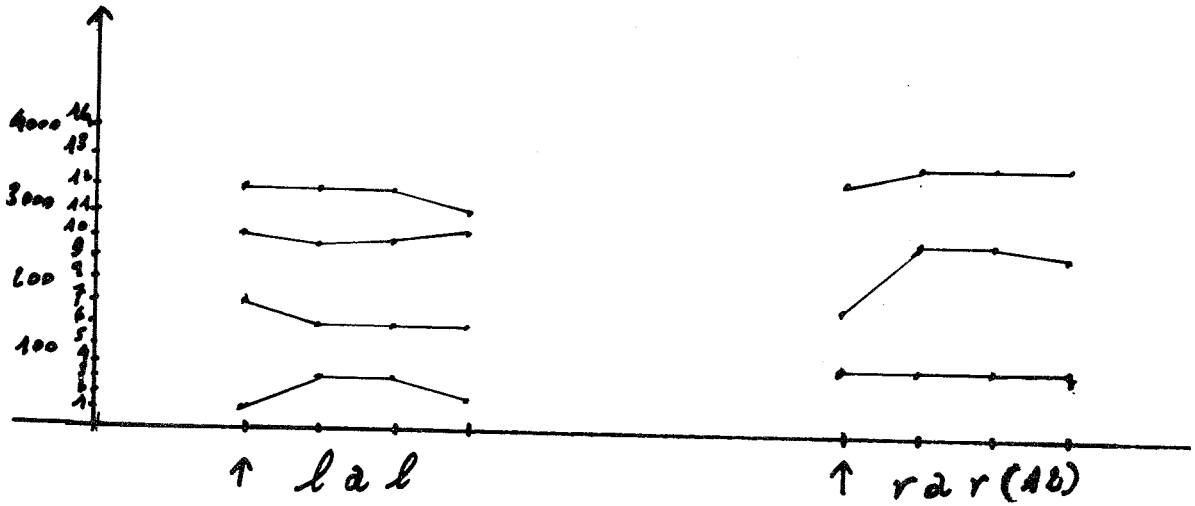


Fig. 7f

Figures 7e et 7f représentent la voyelle /a/ dans les contextes consonantiques nasal et liquide.

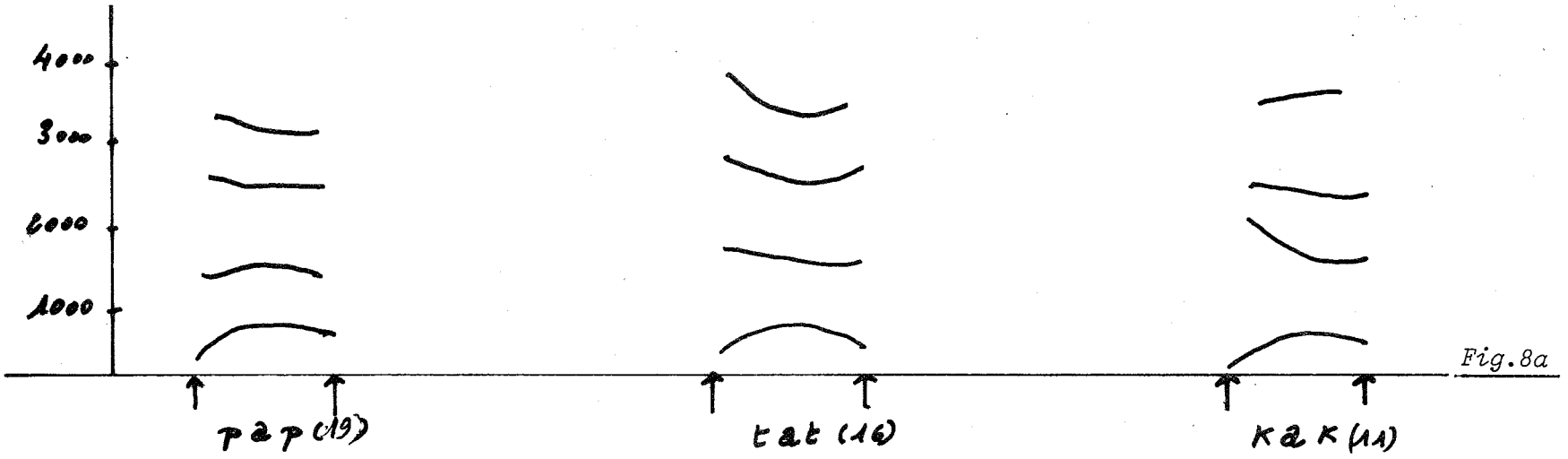


Fig.8a

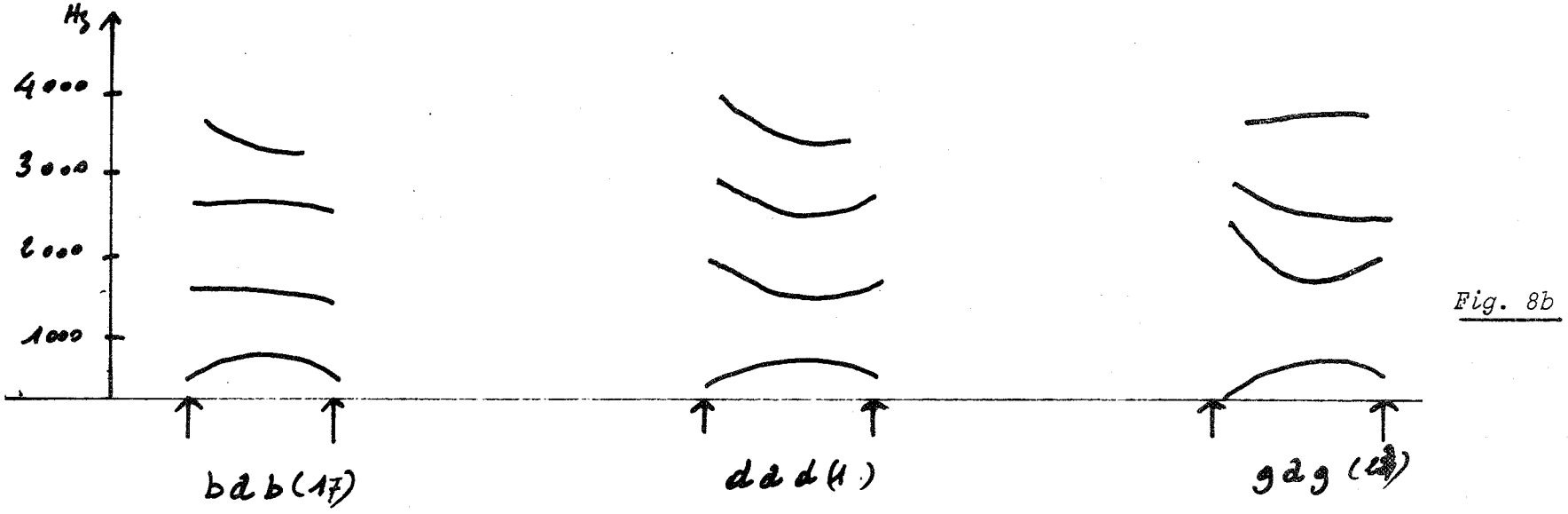


Fig. 8b

Figure 8: montre les formants dans la voyelle /a/ dans les différents environnements consonantiques à partir des données du sonagraphe.

Les figures 8a et 8b représentent la voyelle /a/ dans le contexte consonantique occlusif.

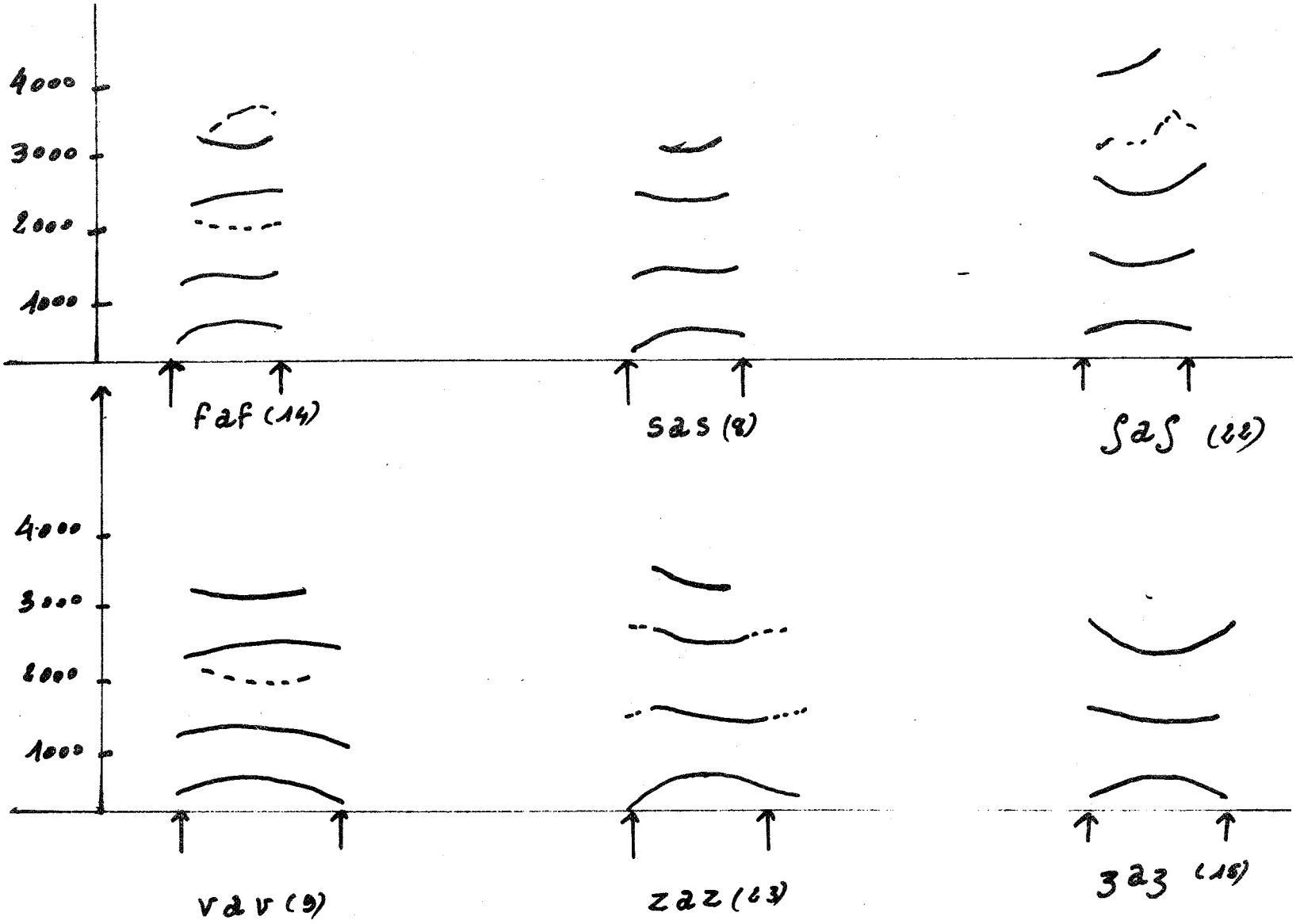
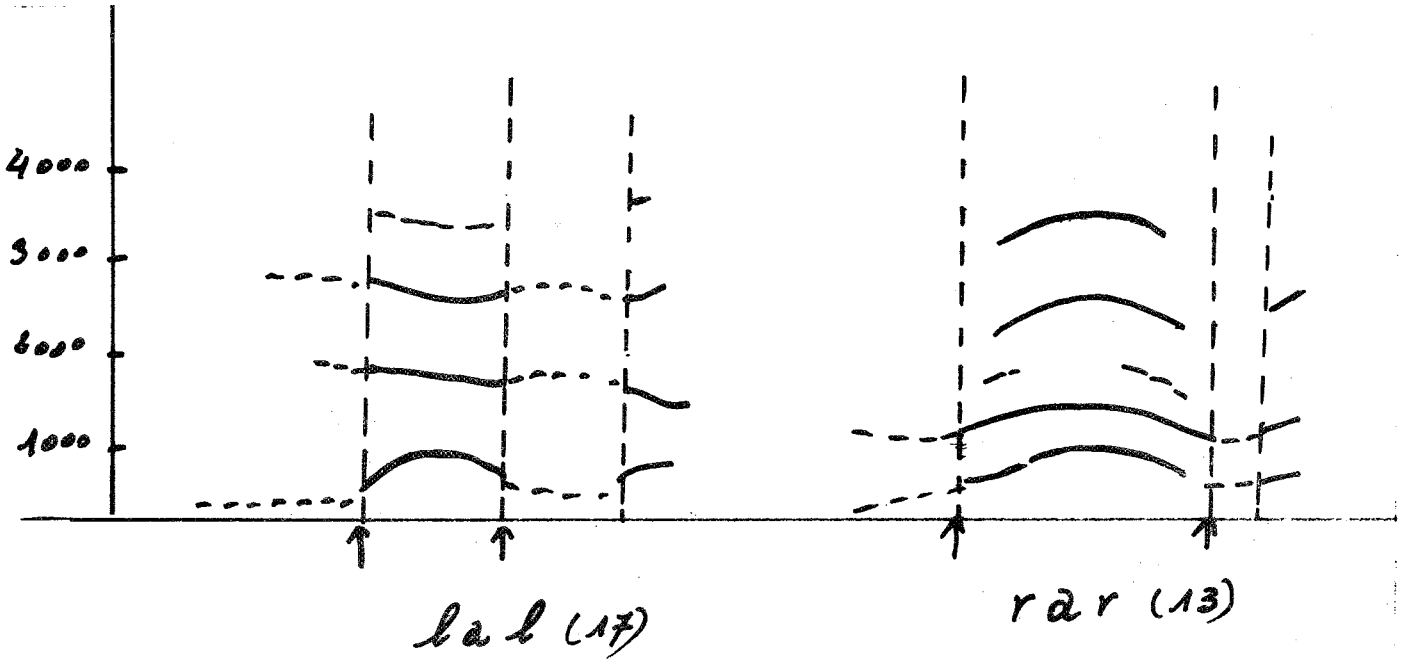


Fig. 8c

Fig. 8d

Les figures 8c et 8d représentent la voyelle /a/ dans le contexte consonantique fricatif.



La figure 8e représente la voyelle /a/ dans le contexte consonantique liquide.

9èmes JOURNEES D'ETUDE SUR LA PAROLE

LANNION 31 mai - 2 juin 1978

LES INDICES DE L'ACCENT TONIQUE ET LEUR HIERARCHIE : APPLICATION A L'ESPAGNOL

G. KONOPCZYNSKI; S. PRUDHAM; R. VARGAS;
C. VITERI (Laboratoire de Phonétique - U.E.R.
Lettres. Université de FRANCHE -- COMTE)
25030 BESANCON Cedex.

RESUME

On étudie les paramètres de l'accent espagnol, qui a une fonction distinctive, dans le cadre du mot. Limitée provisoirement à une analyse de l'onde sonore, l'étude fait apparaître que, des trois paramètres considérés, ni la durée, ni la hauteur ne jouent de rôle appréciable. Les ruptures tonales elles-mêmes ne présentent aucune constante. Seule l'intensité globale, c'est-à-dire la sonie (corrigée selon les caractéristiques intrinsèques des sons) évaluée en fonction de la durée, et non l'intensité maximale instantanée, semble responsable de la proéminence accentuelle réalisée par l'émetteur. Reste à étudier d'une part comment la proéminence est perçue, d'autre part comment les paramètres évoluent dans le cadre de la phrase où interfèrent d'autres facteurs prosodiques.

LES INDICES DE L'ACCENT TONIQUE ET LEUR HIERARCHIE : APPLICATION A L'ESPAGNOL

G. KONOPCZYNSKI; S. PRUDHAM; R. VARGAS
C. VITERI (Laboratoire de Phonétique. U.E.R. Lettres
Université de Franche - Comté), BESANCON

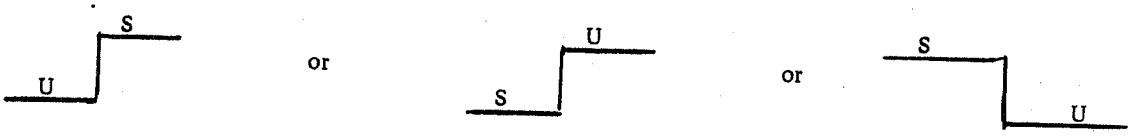
SUMMARY

The purpose of this study is to analyze the features of Spanish stress. In this language, word - stress is linguistically relevant, and its place depends on lexical and morphological rules. But, if the function and place of Spanish stress are well known, its nature is less clear, and there has been much discussion to know which cue - intensity or pitch ? - is the most important. The question has not been proved by enough instrumental evidence.

In our study, 13 groups of 2 or 3 words, only distinguished by their stress, are analyzed by means of spectrographic data. Three main features are examined: duration, fundamental frequency and intensity which is looked at in two ways: maximal intensity and intensity together with duration, by the calculation of the intensity surface in mm^2. Conversion is made in phones and sones, so as to approximate auditory behaviour. A general correction is also introduced to take into account the intrinsic qualities of each vowel.

The study is only concerned with the analysis of the sound wave. We looked for cues which appear only on stressed syllables and not on unstressed ones, and which are, thus, responsible of prominence.

The results (see table 1) show that neither duration nor fundamental frequency are pertinent. The longest vowels are the final ones, not the stressed ones (only 13.6 % of these last are the longest). When the whole syllable is concerned, the percentage goes up to 22.7 %, because the consonants are involved in the stress, but this is still not enough to make duration an essential cue. Looking at the F_0, we can see that tonal breaks are very frequent (80 %), but they are found on stressed syllables (S) as well as on unstressed (U) ones, with the following shape:



The question to ask is: where is the listener supposed to hear the prominence in this case?. Finally, the stressed syllables are not well integrated into the general word-curve, because they are presented in isolation, and not in sentences. Thus we are unable to conclude about this feature.

Finally, only the general intensity has good stability; 72.7 % of the stressed vowels rise higher than the others, even when their maximal intensity is lower than for others vowels (initial ones).

These results must be considered as temporary, since the same words must now be used in sentences whose melodic curve will probably influence the tonal features; we also need to analyze how these physical cues are perceived, especially as pitch is easier to perceive than intensity.

LES INDICES DE L' ACCENT TONIQUE ET LEUR HIERARCHIE: APPLICATION A L'ESPAGNOL

G. KONOPCZYNSKI; S. PRUDHAM; R. VARGAS;
C. VITERI (Laboratoire de Phonétique - U.E.R.
Lettres. Université de FRANCHE - COMTE)
25030 BESANCON-Cedex.

INTRODUCTION

Contrairement au français, dont l'accent dit "logique "(par opposition à l'accent emphatique ou d'insistance dont on ne s'occupera pas ici) occupe une place fixe dans le cadre du mot et du groupe phonique, l'espagnol possède un accent de mot, dont la place, variable (sur les trois dernières syllabes), est déterminée par des règles lexicales ou morphologiques. En outre, alors que la fonction de l' accent français est principalement démarcative, celle de l'espagnol est distinctive, des mots, identiques au niveau phonématique, pouvant avoir un sens différent selon la place de l'accent.

Ex. : inTERprete / interPREte / interpreTE
Trad. : un interprète/ qu'il interprète / j'interprétai
(Nous emploierons l'orthographe usuelle en écrivant la syllabe accentuée en majuscules, pour éviter des confusions entre accent orthographique et accent, symbole employé par l'A.P.I.)

Si la fonction de l'accent en espagnol est bien connue, ainsi que ses règles de distribution, sa nature l'est beaucoup moins.

ETAT DE LA QUESTION

Pour certains auteurs, se fiant à leur seule oreille, l'allongement et / ou une élévation de hauteur, seraient responsables de la proéminence accentuelle. C'est la position présentée en particulier dans la Gramática de la Lengua Española de la Real Academia (1931) ou dans la Gramática de la Lengua Castellana (1949) d'ANDRES BELLO. Pour d'autres, plus rares, une combinaison équilibrée d'intensité et de hauteur (S.F. RAMIREZ - 1951; M. AMADOR - 1953) créerait l'accent, mais aucune analyse instrumentale n'étaie cette opinion.

La plupart des phonéticiens espagnols, reprenant les idées exprimées dès 1918 par NAVARRO TOMAS, dont les Manual de Pronunciacion (1918) et Manual de Entonacion (1944) restent les ouvrages de base, considèrent au contraire, à partir d'études kymographiques, que le paramètre essentiel est l'intensité et optent pour la nature expiratoire et non mélodique de l'accent; les changements de hauteur se manifestant sur la syllabe accentuée seraient de pures coïncidences dues à la courbe mélodique générale de la phrase. Cette position est défendue entre autres par RUFINO CUERVO et plus récemment par ETHEL WALLIS (1951) et par P. DELATTRE (1965: 33) qui, pense que, pour l'espagnol " the main factor of syllabic prominence is a combination of vowel intensity and consonant closure duration".

Récemment, BOLINGER et HODAPP (1961), sur la base de leurs théories accentuelles, s'élevèrent contre cette suprématie accordée à l'intensité. Rejoignant la plupart des phonéticiens qui s'accordent aujourd'hui à dire que le facteur de proéminence accentuelle est la hauteur, et que l'accent ne représente somme toute que la manifestation de l'intonation de phrase au niveau du mot, BOLINGER, après analyses spectrographiques et tests auditifs, conclut à la primauté des ruptures tonales comme paramètre de l'accent espagnol. Mais, outre le fait que les phrases testées contiennent essentiellement des accents d'insistance, et non des accents à fonction distinctive, les résultats restent au niveau qualitatif. Il était donc intéressant de reprendre le problème à sa base.

PROCEDURES EXPERIMENTALES

Etant donné la fonction distinctive de l'accent en espagnol, le corpus est composé de treize couples ou triplets de deux à quatre syllabes, différenciés par la place de l'accent, et présentés en séries à des locuteurs masculins, originaires d'Amérique Latine.

Enregistrement en studio, sur Nagra IV S, avec distance micro - locuteur et niveau d'enregistrement constants.

L'analyse porte sur les paramètres suivants, analysés à l'aide de spectrogrammes:

1) - Durée (en cs.) des voyelles seules et des syllabes. (marge d'erreur≃0.50 cs.).

2) -Fréquence fondamentale de toutes les voyelles, mesurée après l'attaque et en fin d'émission; calcul de la rupture tonale entre les diverses voyelles, en Hz. et en demi-tons. (Marge d'erreur: ± 10 Hz.).

3) - Intensité des voyelles:

a) Valeur instantanée maximale, en dB. ou phones, convertis en sones: Certes, les valeurs initiales sont en dB., non en phones; mais étant donné, d'une part la définition des phones (sonie d'un son pur de 1000Hz., à 40 dB.= 40 phones) et d'autre part la gamme de fréquence la plus importante pour l'intelligibilité du langage (1000 - 2000 Hz.), nous considérons qu'on peut établir, approximativement, une équivalence entre dB. et phones, et de là, convertir en sones, unité physiologique qui présente l'avantage, sur les précédentes, de ne pas être logarithmique et de pouvoir s'additioner et se soustraire, ce qui est important si l'on veut évaluer la chute d'intensité des voyelles. Etant donné l'approximation ainsi introduite, aucune correction d'intensité n'a été préalablement introduite en fonction des valeurs du fondamental.

b) Cependant, le calcul de la sonie instantanée est insuffisant, car la perception de la force sonore dépend étroitement de la durée; pour intégrer ce dernier facteur, nous avons calculé la surface d'intensité de chaque voyelle, en mm^2, selon la technique préconisée par SANTERRE et BOTHOREL (1970). Sans être parfaite - puisque la valeur efficace de l'intensité n'est pas directement prise en compte - cette surface exprime mieux la globalité perceptuelle de la sonie que l'intensité maximale. Il est certes probable que la durée des voyelles atones ($\simeq$ 7 cs.) rende ce calcul illusoire, si l'on suit les conclusions de WOLSK (1964, cité par ROSSI - 1976: 167, "65 %des sujets seulement perçoivent une variation de 50 dB. sur 0.33 sec. !"), mais il est préférable d'avoir la même base de calcul pour les voyelles brèves et longues.

Nous ne tenons pas compte ici de l'intensité spécifique des voyelles, car nous ne disposons pas, pour l'espagnol, d'un tableau de correction semblable à ceux établis par ROSSI pour le français (1971) et l'italien (1976). Pour réduire cette part d'erreur dans nos calculs, huit groupes sur treize contiennent des voyelles identiques donc comparables (BOto - boTO, canTAra -cantaRA....) et pour les autres groupes, les chiffres d'intensité ont été affectés globalement des signes + ou - selon les connaissances théoriques sur les facteurs de pondération en relation avec la structure spectrale des voyelles (ainsi, [i] est affecté de +++ , c'est à dire que son intensité réelle est sensiblement supérieure à notre chiffre, [o] est affecté de + , ...etc.).

Enfin, il faut signaler un artefact dû au corpus. Les mots étant présentés en séries oppositives, celles-ci peuvent être assimilées par les locuteurs à des phrases, et le dernier item du groupe présente alors un contour terminal, avec chute du F_0 et de l'intensité, ce qui fausse les chiffres pour treize voyelles finales; un second corpus présenté de façon différente, remédiera à cet inconvénient.

L'étude consiste à rechercher si un des éléments ci-dessus, de par sa présence constante sur la syllabe accentuée, et son absence sur les autres syllabes, peut être rendu responsable de la proéminence accentuelle réalisée par le locuteur. Les résultats, fondés uniquement, à ce point de nos recherches, sur l'établissement de constantes dans l'onde sonore, négligent donc actuellement le côté psycho-acoustique.

RESULTATS

D'après les données brutes de l'analyse, et chaque paramètre étant considéré isolément, la voyelle accentuée domine les voyelles environnantes par:

- la durée (D) dans seulement 13.6 % des cas.
- la hauteur (F) dans 42.4 % des cas.
- l'intensité maximale instantanée (Im) dans 42.4 % des cas.
- l'intensité globale (Ig) durant la tenue vocalique dans 72.7 % des cas.

En outre, l'accent semble réalisé sans paramètre apparent dans six items, alors que l'on pouvait s'attendre, dans le cadre de mots présentés en opposition, à une insistance sur la proéminence accentuelle.

Les trois paramètres considérés simultanément, il apparaît d'ores et déjà que l'intensité, sans être présente à 100 %, joue un rôle important, puisqu'elle apparaît, à elle seule, dans 33 % des cas, combinée à la hauteur dans 18 % des cas, à la hauteur et à la durée dans 24% des cas, alors que la hauteur n'intervient jamais seule.

cf.: tableau page suivante .

Paramètres des toniques \ ITEMS	Mots de 2 syllabes								Mots de 3 syllabes																						Mot de 4 syllabes		
	1		2		3		4		5		6		7			8			9			10			11			12			13		
	a	b	a	b	a	b	a	b	a	b	a	b	a	b	c	a	b	c	a	b	c	a	b	c	a	b	c	a	b	c	a	b	c
DUREE																																	
- voyelle	-	+	-	+	=	=	-	+	-	=	+	+	-	-	+	-	-	+	-	=	+	-	-	+	-	-	+	=	+	+	-	-	+
- syllabe	-	+	-	+	+	=	n. m.		+	=	+	-	-	-	+	n. m.		=	=	+	+	n. m.			-	-	+	+	-	=	=	-	+
INTENSITE																																	
- maximale instantanée	=	=	+	-	+	-	-	+	+	-	+	=	+	-	-	+	=	-	+	+	-	-	=	=	-	-	=	+	+	+	=	+	=
- corrigée						(++) =				(+++) =													(+++) +				(+) +							
SURFACE																																	
- en mm.	-	+	+	+	+	+	-	+	+	-	+	+	-	=	+	-	+	+	-	+	+	+	+	+	=	-	+	+	+	+	+	=	+
- corrigée										(+++) =									(+++) -															
F_o	-	+	+	-	-	+	-	+	-	-	+	-	+	-	-	-	+	+	+	+	-	-	-	+	-	=	+	+	+	+	-	=	-

TABLEAU 1 Les paramètres des voyelles toniques
Features of stressed vowels

COMMENTAIRE: Le signe + indique que la voyelle tonique domine les autres voyelles en ce qui concerne le trait étudié
- Dans les mots a la syllabe initiale porte l'accent (sauf dans treize cas où c'est la seconde syllabe)
- Dans les mots b des groupes 1 - 4 et c des groupes 7 - 13 les syllabes finales portent l'accent
- Dans les mots b des groupes 5 à 13 les syllabes médianes portent l'accent
- n. m. = non mesurable/

En étudiant dans le détail chacun des paramètres, il appert:

1. - DUREE

Les voyelles finales sont quasiment toujours les plus longues, qu'elles soient accentuées ou non. La différence de durée entre les toniques et les finales non toniques, entre 2.70 et 12 cs.,(M ≃ 5 cs.) représente une marge confortable au-dessus du seuil différentiel pour lequel nous retenons la valeur de 2 cs. pour les voyelles les plus brèves et un pourcentage d'augmentation de 22.5 % pour les autres(ROSSI 1972). Il résulte de cet allongement généralisé des finales que les finales toniques sont particulièrement longues: 18.5 cs. en moyenne / finales atones ≃ 14 cs.

En seconde place après les finales viennent les toniques, (≃ 10.5 cs.); les atones, pré-ou posttoniques, durant ≃ 7.5 cs.

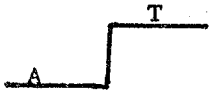
Si l'on prend en considération les syllabes entières, on constate que, dans 22.7 % des cas, la syllabe tonique (M: ≃ 19.5 cs.) domine la finale, alors qu'au niveau des voyelles, ce pourcentage s'abaisse à 13.6 %. Le comportement des consonnes, dont l'augmentation quantitative certaine sous l'effet de l'accent (P. SIMON, 1967: 160) est à incriminer, n'a cependant pas été analysé de près, le corpus n'étant pas équilibré de ce point de vue (on ne peut pas comparer les durées des consonnes nasales et des occlusives sourdes); dans un second corpus, en cours d'analyse, le choix des consonnes permettra une étude de ce point précis. Cependant, même dans le cadre de la syllabe, le paramètre durée ne peut être considéré comme constant, et, considéré isolément, il ne peut être un indice d'accent. Mais le problème est à revoir en fonction de l'intensité, la durée ne pouvant, de l'avis général, être regardée comme un paramètre prosodique indépendant.

2. - FREQUENCE FONDAMENTALE

La tonique domine les autres voyelles par un accroissement de hauteur dans 42.4 % des cas; contrairement à la durée, aucune autre position ne présente de constante pour ce paramètre, les voyelles les plus hautes étant indifféremment les pré - ou les posttoniques, les initiales ou les finales. Notons cependant que le contour terminal dont sont pourvues les treize voyelles finales de chaque série n'affecte guère leur hauteur, puisque le petit groupe des toniques dominant par la hauteur, est constitué presque à 50 % de finales accentuées.

Si l'accent n'est pas réalisé par une hauteur plus importante des toniques, il ne l'est pas non plus par leur contour mélodique. En effet, mises à part les treize voyelles finales de séries déjà évoquées, au contour descendant avec un F_0 terminal plus bas que le reste du mot, les autres ne présentent aucune constante, et se répartissent environ par moitié, les unes montantes, les autres descendantes, quelques rares voyelles ayant un contour plat.

La voyelle tonique ne se caractérise donc ni par un sommet de hauteur, ni par un contour ascendant. Reste à considérer la rupture fréquentielle que BOLINGER (1961) estime essentielle dans le cas de l'accent en général et de celui de l'espagnol en particulier. Dans le cas où la tonique (T) est précédée de voyelle(s) atone(s) (A), et suivie ou non des voyelles atones, 81 % des cas présentent une rupture de forme:



; le décalage vers le haut de la tonique (compris entre 13 et 57 Hz. soit, pour le niveau de F_0 considéré, de 1 demi-ton à 6 demi-tons; M ≃ 33 Hz., donc largement supérieur au seuil) pourrait donc être un indice de l'accent, du moins dans cette position. En revanche pour les toniques en initiale, la fréquence d'apparition est identique pour la forme:



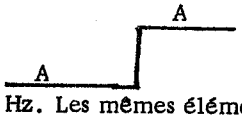
et la forme:



Dans ces conditions, où l'auditeur entendra-t-il l'accent?

Un examen précis des cas impliqués montre par ailleurs que la moitié d'entre eux se réalise sans paramètre apparent, et l'autre moitié avec un maximum de sonie.

Enfin le rôle de la rupture fréquentielle, présente dans un nombre important de cas, reste d'autant plus difficile à déterminer que les voyelles atones présentent les mêmes caractéristiques; deux voyelles atones successives ont également le contour:



ou



, avec une différence allant de 5 à 34 Hz., M ≃ 20 Hz. Les mêmes éléments frappant aussi bien voyelles accentuées qu'inaccentuées ne peuvent être considérés comme un indice de l'accent. Nous n'avons pas calculé les variations du fondamental en fonction de la durée; en effet ces glissandos n'auraient pas apporté de nouvelles informations, durée et hauteur n'étant pas un paramètre constant caractéristique des seules toniques. Cependant ce facteur sera pris en compte lorsque les mots sont insérés dans des phrases.

A remarquer également que les voyelles s'intègrent mal dans une courbe mélodique générale, car les mots ne sont pas insérés dans des phrases; on ne peut donc, à ce point, conclure définitivement quant à l'importance des ruptures tonales.

3. - INTENSITE

Les résultats tiennent compte des corrections globales d'intensité; le seuil différentiel retenu est celui indiqué par ROSSI (1976: 168), soit 2 dB.

Si l'on tient compte de l'intensité maximale instantanée, la voyelle tonique domine les autres dans 54.5 % des cas, la voyelle initiale dans 27 % des cas. Les finales, qui l'emportaient du point de vue temporel, ne sont jamais les plus intenses, règle d'ailleurs générale: "l'intensité vocalique des voyelles en position accentuée est plus faible pour la voyelle en finale absolue; la consonne qui suit la voyelle lorsque celle-ci est fermée, a une influence sur la voyelle qui précède (MUNCH, 1970: 118) . Or notre corpus ne comporte que des syllabes finales ouvertes.

Si l'on considère l'intensité globale (Ig), les chiffres deviennent plus significatifs, puisque la tonique domine dans 72.7% des cas; en revanche les initiales se voient reléguées au second plan par les finales qui dominent dans 27 % des cas en raison de leur durée qui est prise en compte dans le calcul de l'Ig. De plus, lorsque une atone domine la tonique, la différence est assez minime comparée à la nette dominance des toniques par rapport aux atones.

De surcroît, notre technique de calcul surestime légèrement, du point de vue qualitatif (et non quantitatif), l'intensité globale des atones. En effet, les voyelles toniques présentent une intensité à peu près stable alors que les atones, et plus particulièrement les finales accusent sur leur dernier quart (ou tiers) des chutes d'intensité importantes, quoique graduelles. Plus parlante que l'intensité globale serait donc, soit une indication de la sonie en procédant à une intégration des diverses intensités (toutes les 2 cs.) soit plutôt un calcul de la chute d' intensité . En raison de l'artefact introduit dans notre corpus par la présentation des items en séries, nous n'avons pas jugé utile de procéder à ces corrections, d'autant plus que, d'après ROSSI (1976 : 167) "l'intensité sonore est donnée par le niveau sonore sur le premier tiers de la voyelle". Malgré cette surestimation il apparaît que la sonie globale, paramètre le plus constamment présent sur les voyelles toniques et généralement absent sur les atones, rend le mieux compte du relief accentuel. Les impressions subjectives des auteurs, qui, percevant nécessairement l'intensité en fonction de la durée, optèrent pour l'intensité, se voient ainsi confirmées par l'analyse instrumentale, et justifient, si besoin était, la nécessité d'une étude globale de l'intensité, au détriment de la seule intensité maximale.

CONCLUSIONS ET PERSPECTIVES

Il semble que l'on puisse établir provisoirement une hiérarchie des paramètres de l'accent espagnol, dans le cadre du mot, au niveau de sa production. La durée en elle-même, ainsi que la hauteur ou le contour mélodique de la voyelle n'interviennent pratiquement pas; le rôle des ruptures tonales, presque toujours présentes, reste difficile à déterminer avec précision, car les mêmes ruptures se manifestent sur les voyelles atones.

En revanche, l'intensité globale, de par sa présence qualitative et quantitative, peut être considérée comme l'indice essentiel de l'accent.

Il reste à établir des constantes au niveau perceptuel; si la sonie apparaît d'ores et déjà un paramètre primordial à ce niveau, on ne peut à priori nier l'existence d'une relation entre certains types de ruptures tonales et la perception de la proéminence.

Enfin, le corpus, élargi (mots isolés, et mêmes mots insérés dans des phrases de structure identique) et réalisé dans des conditions différentes (choix aléatoire fait par les locuteurs) pour éviter les effets de la présentation en série est actuellement en cours d'analyse; on se propose d'étudier l'évolution des paramètres de l'accent dans le cadre de la phrase, avec sa ligne tonale déterminée. La sonie reste-t-elle déterminante, ou bien les ruptures tonales prennent - elles le pas , du moins au niveau perceptuel où l'efficacité du paramètre hauteur par rapport à la sonie n'est plus à démontrer?

BIBLIOGRAPHIE

BOLINGER D. L. - HODAPP M. (1961), "Acento melodico, acento de intensidad", Boletin de Filologia de Chile, Vol. 13, pp.33 - 48.

DELATTRE P. (1965), Comparing the phonetic features of English, German, Spanish and French. Heidelberg: Julius Groos Verlag.

MUNCH JOSIANE (1970), "Quelques observations sur l'intensité des voyelles du français". in Travaux de l' Institut de Phonétique de Strasbourg, Vol. 2, pp.:112 - 125.

NAVARRO TOMAS T.(1918), Manual de Pronunciacion Española. Edition consultée: 19ème édition. Madrid, Consejo Superior de Investigaciones Cientificas. (1977).

NAVARRO TOMAS T.(1944), Manual de Entonacion Espanola. New York.

ROSSI M. (1971), "L'intensité spécifique des voyelles". Phonetica, Vol. 24/3, pp.129 - 161.

ROSSI M. (1972), "Le seuil différentiel de durée" in Papers to the Memory of PIERRE DELATTRE, La Haye: Mouton; pp.:435 - 450.

ROSSI M. (1976), Contribution à la méthodologie de l'analyse linguistique avec application à la description phonétique et phonologique du parler de Rossano. Paris: Champion.

SANTERRE L. , BOTHOREL A. (1970), "Mesures et interprétation de la ligne d'intensité sur les sonagrammes". Travaux de l'Institut de Phonétique de Strasbourg, Vol.: 2, pp.:82 - 90.

SIMON P. (1967), Les consonnes françaises; mouvements et positions articulatoires à la lumière de la radiocinématographie. Paris : Klincksieck.

WALLIS E. "Intonational stress patterns of contemporary spanish", Hispania: 34, pp.: 143 - 146.

ANNEXE

Composition du corpus

Dans le tableau 1, a réfère au premier mot, b au second, c au troisième.

	a	b	c
1. -	BOto	boTO	
2. -	MAma	maMA	
3. -	VIAje	viaJE	
4. -	CANto	canTO	
5. -	Avido	haBIdo	
6. -	SAbana	saBAna	
7. -	CElebre	ceLEbre	celeBRE
8. -	CANtara	canTAra	cantaRA
9. -	CIRculo	cirCUlo	circuLO
10. -	PRACtico	pracTIco	practiCO
11. -	VOmito	voMIto	vomiTO
12. -	GEnero	geNEro	geneRO
13. -	inTERprete	interPREte	interpreTE

9èmes JOURNEES D'ETUDE SUR LA PAROLE

LANNION 31 mai - 2 juin 1978

MODIFICATIONS INTER-LOCUTEURS DE L'ECHELLE FORMANTIQUE

S. BARTH R. CHULLIAT

INSTITUT NATIONAL DE JEUNES SOURDS DE CHAMBERY

CENTRE D'AUDIOLOGIE ET D'ACOUSTIQUE

BP 15 73160-COGNIN

.RESUME

La loi de transposabilité (VON EHRENFELS 1890) exprime le fait que certaines propriétés d'une forme ne sont pas altérées par des transformations (anamorphoses) affectant d'une manière comparable toutes ses parties.

A partir de la mesure des formants vocaliques pratiquée sur 32 sujets, nous discuterons de la validité d'un modèle de modification inter-locuteurs de l'échelle formantique.

Les résultats présentés sont en bonne concordance avec les données de la physiologie de la phonation et montrent l'intérêt de l'application d'une transformation non linéaire de l'échelle formantique au problème de l'adaptation au locuteur.

SUMMARY

VON EHRENFELS's law states that certain characteristics of a Gestalt are not changed by the distorsions affecting in a similar way each of its parts.

Starting from the measurements of vocal formants made on 32 talkers, we shall discuss the validity of a model of a non-linear transformation of the formants frequency scale between the vowels of two speakers.

The results presented are in close agreement with the data of the physiology of speech production and shows the advantage of the application of a non-linear transformation of the formants frequency scale to the problem of between speakers adaptation.

I - INTRODUCTION

Les fréquences formantiques des voyelles ont fait l'objet de nombreux relevés. Les plus anciens d'entre eux (BARNEY et PETERSON, 1952) mettaient déjà en évidence une dispersion inter-individuelle de ces paramètres se manifestant par des formants plus élevés pour des enfants que pour des hommes ou des femmes adultes, dispersion plus nettement marquée entre les hommes et les enfants qu'entre ceux-ci et les femmes.

Parmi les causes possibles de ces fluctuations, il semble naturel de retenir la variation inter-individuelle de la longueur intrinsèque du conduit vocal (CV). En effet, les formants sont les pôles de la fonction de transfert du CV dont la configuration est supposée stationnaire pendant un temps suffisamment long. La valeur de ces pôles est, de ce fait, étroitement liée aux caractéristiques géométriques du CV.

Si l'on se rapporte aux mesures publiées par CHIBA et KAJIYAMA (1941), la longueur moyenne du CV d'une femme adulte est de l'ordre de O.87 lorsque celle d'un homme adulte est prise comme unité. On peut donc s'attendre, sur la base du [ə], dont le modèle acoustique est un tube uniforme, à des formants plus élevés pour les femmes dans un rapport égal à la diminution relative de la longueur du CV : soit 13%.

MOL (1963), reprenant les mesures de BARNEY et PETERSON, rechercha une variation linéaire des fréquences formantiques avec l'âge et posa le principe de la "croissance axiale uniforme" du CV. Les résultats qu'il obtint se basaient malheureusement sur un nombre trop restreint de sujets et les relevés publiés plus tard par EGUSHI et HIRSH (1969) allaient en montrer les limites.

FANT (1966) affirma à juste titre qu'il existait d'autres différences anatomiques que la longueur intrinsèque entre les CV des hommes et des femmes. Ceux des enfants ressemblant anatomiquement plus à ceux des femmes, il se produit, lorsqu'un garçon grandit, des variations des fréquences formantiques de ses voyelles qui ne peuvent pas être assimilées à celles que l'on observe lorsque l'on augmente la longueur d'un tube uniforme.

E. LEIPP (1968) partant des données de la Gestaltthéorie formula un modèle de perception susceptible de permettre la description de telles variations : l'information phonétique est véhiculée par le squelette sémantique qui, en tant que Super-Forme, peut subir certaines altérations suivant l'axe des fréquences ou (et) celui du temps sans que son intelligibilité en soit perturbée.

C'est à une forme générale de distortion fréquentielle du squelette sémantique que nous nous intéresserons dans les lignes qui suivent.

II - L'ANAMORPHOSE FREQUENTIELLE IRREGULIERE

Si l'on note $x2(t,F)$ une Forme de référence du plan temps-fréquence et $x1(t,F)$ son anamorphosée, l'anamorphose fréquentielle irrégulière peut se définir, selon J.S. LIENARD, par la transformation :

$$\begin{cases} x1(t,F) = x2(t, u(F)) \\ u(F) = KF + E(F) \\ K > O \end{cases}$$

Nous verrons plus loin que le choix d'une fonction constante pour E (F) assure une bonne approximation de u (F).

L'objet de cette étude portant sur les fréquences formantiques, nous négligerons, en première approximation le paramètre t en mesurant celles-ci aux instants de plus forte stabilité de la voyelle des groupes CV : vi, ve, vɛ , vy, voe, vu, vo, va.

Compte-tenu du type de relation à chercher et de la symétrie possible des rôles des locuteurs "donnée" et "référence", nous avons opté pour un modèle de regression orthogonale défini comme suit :

soit F' et F les vecteurs colonnes de composantes f'_i et f_i (où f_i et f'_i sont les fréquences formantiques pour la référence et la donnée de l'ensemble du système vocalique). La droite de regression orthogonale est fournie par :

$$\begin{cases} F' = KF + b \\ K = \dfrac{2\hat{r}}{S_o + \sqrt{S_o^2 + 4\hat{r}^2}} \\ b = \overline{F'} - K\overline{F} \\ \hat{r} = \dfrac{\mathrm{cov}(F', F)}{\sqrt{\mathrm{var}\, F \;\; \mathrm{var}\, F'}} \\ \overline{F'} = \dfrac{1}{n}\sum_i f'_i \;, \quad \overline{F} = \dfrac{1}{n}\sum_i f_i \\ S_o = \dfrac{\sigma_F}{\sigma_{F'}} - \dfrac{\sigma_{F'}}{\sigma_F} \;, \quad \sigma_F^2 = \dfrac{1}{n}\sum_i (f_i - \overline{F})^2, \quad \sigma_{F'}^2 = \dfrac{1}{n}\sum_i (f'_i - \overline{F'})^2 \end{cases}$$

Cette droite ajuste le nuage des points (f_i, f'_i) au sens des moindres carrés, les résidus étant calculés orthogonalement à celle-ci sans favoriser l'une ou l'autre des directions. Cette formulation permet d'avoir facilement la relation inverse :

$$\begin{cases} F = K' F' + b' \\ K' = \dfrac{1}{K} \\ b' = -\dfrac{b}{K} \end{cases}$$

III - MODALITES D'EXPERIMENTATION

Nous disposions d'un corpus de 32 locuteurs (5 hommes adultes, 5 femmes adultes, 6 garçons et 6 filles de 9 ans, 5 garçons et 5 filles de 6 ans). Pour chacun d'eux, nous avons mesuré les fréquences formantiques (F1, F2) des voyelles i, e, y, ɛ , u, o, oe, ɑ enregistrées en chambre sourde sur magnétophone REVOX A77 (bande SONY HL, micro SENNHEISER MD 441). Les formants ont été assimilés à des maxima spectraux et, pour les cas difficiles où F1 coïncidait avec F_o, nous nous sommes servis des propriétés du spectre différentiel (J.S. LIENARD, 1975) comme critère de décision. L'analyse spectrale a été pratiquée par des programmes, développés en T.S.L. pour l'ensemble PDP 11/20 - analyseur TIME DATE 1923 (GENRAD) du laboratoire, donnant un $\Delta F \simeq 300$ HZ et une précision de $\pm$ 40 HZ.

L'ensemble des données recueillies a été traité par un programme écrit en FORTRAN IV donnant, pour chaque couple de locuteurs :

1) les coefficients K et b définis au § II
2) la variance extraite en % de la variance totale
3) l'intervalle de confiance pour K donné par :

$$\hat{K} - t_{\alpha/2}\,(n-2)\,\frac{S}{\sqrt{n}\,\sigma_F} < K < \hat{K} + t_{\alpha/2}\,(n-2)\,\frac{S}{\sqrt{n}\,\sigma_F}$$

où $t_{\alpha/2}\,(n-2)$ est la valeur de la distribution de STUDENT à n - 2 degrés de liberté ayant la probabilité α d'être dépassée, $\hat{K}$ l'estimation de K calculée, n le nombre de fréquences formantiques entrant dans le calcul, σ_F l'écart type de F et S l'estimation de la variance résiduelle $S = \frac{1}{n-2} \sum_i (f'_i - Kf_i - b)^2$

4) le coefficient de correlation de rang de SPEARMAN :

$$\rho_s = 1 - \frac{\sum d_i^2}{n^3 - n} \quad \text{avec } di = \text{rang } (f'_i) - \text{rang } (f_i)$$

IV - RESULTATS

IV.1 Test sur ρ_s

	ADULTES	9 ANS	6 ANS
ADULTES	1.00 0.95	0.90 0.85	0.87 0.83
9 ANS		1.00 0.96	0.92 0.85
6 ANS			1.00 0.94

Le tableau ci-dessus donne les valeurs maxima et minima de ρ_s calculé dans chaque groupe (sur la diagonale) puis en associant les groupes entre eux. La comparaison de ces valeurs à $t_{\alpha/2}\,(n-2)\sqrt{\frac{1-\rho_s^2}{n-2}}$ permet de vérifier qu'elles sont toutes significatives d'une liaison au seuil α = 10%.

IV.2 Pour les adultes

		Δ K	variance extraite
H1	F' = F	0.00	100
H2	F' = 1.12F - 101	0.05	96.08
H3	F' = 1.11F	0.05	98.66
H4	F' = 1.05F	0.04	99.02
H5	F' = 1.18F - 66	0.04	99.20
F1	F' = 1.25F - 47	0.05	96.53
F2	F' = 1.37F - 53	0.06	97.13
F3	F' = 1.20F	0.07	98.20

F4	F' = 1.22F - 65	O.O5	98.22
F5	F' = 1.2OF	O.O7	98.15
	Locuteur de référence : H1		
	Locuteurs masculins : H1, H2, ..., H5		
	Locuteurs féminins : F1, F2, ... F5		

IV.3 Pour les enfants de 9 ans

		Δ K	variance extraite
F 1	F' = F	O.OO	1OO
F 2	F' = 1.O2F	O.O3	96.34
E 3	F' = O.96F + 89	O.O3	99.55
F4	F' = 1.O1F + 3O	O.O4	97.42
F5	F' = O.99F	O.O4	98.22
F6	F' = O.99F	O.O3	99.42
G1	F' = O .95F	O.O3	99.58
G2	F' = 1.O5F - 42	O.O6	98.OO
G3	F' = 1.O2 F	O.O3	9O.82
G4	F' = O.96F	O.O2	99.9O
G5	F' = O.94F	O.O3	99.51
G6	F' = O.99F - 32	O.O4	96.O4
	Locuteur référence : F1		
	Locuteurs masculins : G1, G2, ..., G6		
	Locuteurs féminins : F 1, F2, ..., F6		

IV.4 Pour les enfants de 6 ans

		Δ K	Variance extraite
G1	F' = F	O.OO	1OO
G2	F' = 1.13F	O.O3	99.33
G3	F' = 1.O9 F	O.O4	99.72
G4	F' = 1.15F	O.O5	97.87
G5	F' = 1.19F - 112	O.O5	8O.69
F1	F' = 1.O2F - 76	O.O3	96.42
F2	F' = 1.O5F - 1O2	O.O4	99.73
F3	F' = O.99F	O.O2	98.26
F4	F' = O.98F - 172	O.O3	99.43
F5	F' = 1.19F	O.O5	97.75
	Locuteur de référence : G1		
	Locuteurs masculins : G1, G2, ..., G5		
	Locuteurs féminins : F1, F2, ..., F5		

IV.5 Cas où H1 sert de référence aux enfants

			Δ K	variance extraite
9 ANS	F1	F' = 1.66F - 145	O.28	83.88
	F2	F' = 1.73F - 145	O.37	71.82
	F3	F' = 1.59F - 55	O.29	79.78
	F4	F' = 1.71F - 147	O.35	73.68
	F5	F' = 1.67F - 145	O.33	76.5O
	F6	F' = 1.65F - 145	O.3O	8O.2O
	G1	F' = 1.58F - 145	O.28	81.32
	G2	F' = 1.77F - 218	O.34	77.75
	G3	F' = 1.63F - 145	O.11	97.37
	G4	F' = 1.6OF - 145	O.28	81.6O
	G5	F' = 1.56F - 145	O.28	8O.52
	G6	F' = 1.68F - 2O1	O.34	77.51
6 ANS	G1	F' = 1.71F - 133	O.27	79.24
	G2	F' = 1.7OF - 133	O.32	76.5O
	G3	F' = 1.64F - 133	O.34	74.25
	G4	F' = 1.73F - 133	O.34	74.1O
	G5	F' = 1.79F - 245	O.36	71.O2
	F1	F' = 1.53F - 2O9	O.29	88.26
	F2	F' = 1.58F - 235	O.29	87.9O
	F3	F' = 1.49F - 133	O.25	89.2O
	F4	F' = 1.47F - 3O5	O.23	89.OO
	F5	F' = 1.79F - 133	O.35	75.1O

V - DISCUSSION

V.1 Valeur moyenne de K

Lorsque H1 sert de référence, on obtient :

	K moyen	σ(K)
HOMMES ADULTES	1.O9	O.O4
FEMMES ADULTES	1.24	O.O6
FILLES	1.62	O.O9
GARÇONS	1.67	O.O7

La différence est significative au seuil de O.9O entre les hommes et les femmes adultes alors qu'elle ne l'est pas entre les garçons et les filles (6 ans et 9 ans).

Il faut noter que le rapport des coefficients K moyens des hommes et des femmes adultes est de l'ordre de 1.13. Ce résultat est à mettre en parallèle avec l'augmentation de 13% des fréquences formantiques prévisible à partir des mesures de CHIBA et KAJIYAMA.

V.2 Cas d'anamorphose régulière

On observe des cas d'anamorphose régulière ($b \neq 0$) lorsque le locuteur

"de référence" est choisi dans le même groupe que les locuteurs "données". Par contre, ils ne peuvent être mis en évidence par panachage des groupes que dans 7% des cas, correspondant à la comparaison d'enfants avec des locuteurs féminins. Ce résultat tend à infirmer l'hypothèse de la "croissance axiale uniforme" du CV. Il concorde, d'une manière moins nette, avec celle de la ressemblance anatomique des CV de femmes et d'enfants.

VI - CONCLUSION

L'anamorphose fréquentielle définie au § II permet de trouver, d'une manière simple et efficace, la modification de l'échelle formantique entre deux locuteurs. Les résultats que nous avons présentés concordent avec les données de la physiologie de la phonation. Ils ne sont bien sûr qu'une approximation car la valeur des coefficients K et b dépend aussi du contexte phonétique, donc du temps.

BIBLIOGRAPHIE

S. BARTH, R. CHULLIAT : Etude sur l'anamorphose des voyelles en fonction de l'âge. Bulletin d'Audiophonologie de BESANÇON, vol I (1978)

S. EGUSHI et I.J. HIRSH : Development of speech sounds in children. Acta Otolaryng. Suppl 257 (1969)

G. FANT : A note on vocal tract size factors and non-uniform F. pattern scaling. Quat.Rep. R.I.T. n° 4, 22-30 (1966)

H. MOL : Fundamentals of Phonetics. Janua Linguanum n° 26, La Hague, Mouton (1963)

J.S. LIENARD : Processus de la communication parlée. Masson (1977)

J.S. LIENARD : Differential analysis of speech. 8e Congrès de Phonétique LEEDS (1975)

E. LEIPP : Information sémantique et parole - GAM n° 22, PARIS VI (1966)

9èmes JOURNEES D'ETUDE SUR LA PAROLE

LANNION 31 mai - 2 juin 1978

ESSAI D'ANALYSE PHONOLOGIQUE DES INDICES DU VOISEMENT

Christian ABRY *
Louis-Jean BOE

Institut de Phonétique de Grenoble

.**RESUME**

L'analyse phonologique classique de distinctivité opère sur des traits. Cette étude est un essai d'adaptation de ces méthodes aux indices dégagés depuis plusieurs années par les recherches phonétiques sur le voisement. L'analyse menée sur sept indices pour les occlusives en français, permet d'établir, à partir de leurs simples relations logiques, une échelle de distinctivité des indices du voisement, qui n'est pas sans rapport avec les résultats que nous ont fourni jusqu'à présent les analyses perceptives. Ainsi, nous pensons que les hypothèses phonologiques sur les indices, si elle restent bien entendu en attente de l'expérimentation psychophonétique, peuvent servir à élaborer ses stratégies.

* Nous tenons à remercier D. CREISSELS (Linguistique, Grenoble III), qui a bien voulu discuter la première version de cet exposé.

PHONOLOGICAL ANALYSIS OF VOICING CUES IN FRENCH PLOSIVES

Christian ABRY
Louis-Jean BOË
Institut de Phonétique de Grenoble

. SUMMARY

Phonological analysis of distinctiveness classically operates on features. This paper is an attempt to adapt the same procedures to the cues, in our case, the voicing ones. Among those which have been investigated in the past years of phonetic research, we selected the following seven : Dvoc (preconsonantal vowel duration), Dcons (closure duration), Dprevois (prevoicing or negative VOT), I (intensity - duration of the burst), Dasp (duration of the aspiration or positive VOT, the burst excepted), T1 (first formant transition), Fo (vocalic voicing onset pitch or pitch skip/dip). They are all present for French plosives for which we made such an analysis to evidence logical relation between the cues. So we obtain a scale of distinctiveness which can be well related with the results presently available concerning the perceptual weight of the same voicing cues : the first four being T1, Dasp, Dcons and Dvoc, followed by Dprevois and Fo (perceptually poorly decisive), the last being I (the less tested one). These results make it desirable to conduct such a phonological analysis on phonetic data before setting up experimental strategies, which are finally the only means to test phonological hypothesis.

ESSAI D'ANALYSE PHONOLOGIQUE DES INDICES DU VOISEMENT

Christian ABRY
Louis-Jean BOË
Institut de Phonétique de Grenoble

INTRODUCTION

L'analyse de distinctivité de la phonologie classique se présente comme une recherche, dans la matrice des traits (disons binaires, pour simplifier), des relations logiques : implication, incompatibilité etc... Cette analyse peut s'appliquer à des matrices de :

(1) *traits / phonèmes* où les entrées colonnes représentent l'inventaire des phonèmes,
(2) *traits / allophones (variantes de phonèmes)* où les colonnes sont constituées par les variantes libres ou conditionnées d'un ou plusieurs phonèmes,
(3) *traits / suite de phonèmes* où les colonnes sont constituées par les séquences de phonèmes d'un morphème où d'une suite de morphèmes.

Toutes ces analyses ont un but commun : éliminer les redondances de ces matrices. Ainsi une relation d'incompatibilité entre deux traits permet de n'en garder qu'un, une relation d'implication permet de ne pas spécifier dans les matrices la valeur du trait impliqué (redondant).

Par exemple, dans une matrice de type (1), le voisement est généralement considéré comme redondant pour les colonnes des nasales, puisque *nasal* présuppose *voisé* dans la définition systématique des phonèmes.

Une même analyse à un autre niveau est celle des matrices de type (2). Elle permet de mettre en évidence, par exemple, que, dans les matrices de variantes de nasales, le trait *voisé* n'est pas une constante (on peut avoir des nasales voisées ou non selon le contexte); il peut alors être considéré aussi comme redondant (une analyse classique de ce type est celle de *voisé* et *tendu* en français).

L'analyse des matrices de traits articulatoires qui montre que, des deux traits *occlusif* et *bilabial* de /p/, seul *bilabial* implique *occlusif*, se fait au niveau d'une matrice de type (1) <u>après</u> que les matrices (2) aient montré que ces deux traits étaient indissolublement des constantes de /p/.

ANALYSE DES INDICES DU VOISEMENT

Nous allons aborder l'analyse de distinctivité des indices du voisement par une matrice *indices / variantes de traits (allotraits)* constituée avec sept indices : les 6 de DELATTRE (1967), plus le "pitch skip" (HAGGARD & al., 1970), qui ont sucité au minimum quelques travaux. Nous nous sommes limités aux occlusives, en ne considérant parmi les transitions que celles de F1 de la voyelle postconsonantique et en ne retenant des variations de F_0 que celles de l'établissement du voisement vocalique (nous éliminons celles de la consonne). Ces indices sont :

D_{voc} : la durée de la voyelle préconsonantique,
D_{cons} : la durée de la tenue de l'occlusive,
$D_{prévois}$: la durée du prévoisement (V.O.T. négatif),

I : l'intensité (et la durée) de l'explosion de l'occlusive,

D_{asp} : durée de l'aspiration (V.O.T. positif = D.E.V., moins durée de l'explosion),

T_1 : transition du premier formant F1 de la voyelle postconsonantique,

F_o : mode d'établissement du voisement vocalique (pitch skip).

Nous illustrons 5 positions présentant 11 combinaisons différentes des indices du voisement (11 allotraits), tous exemples qui se rencontrent en français avec des assimilations de sonorité et / ou de détente (Tableau I). Un + dans une case indique que l'indice peut fonctionner (indice différentiel : l'autre membre de la paire est différent sur ce point), un - signifie qu'il est neutralisé.

position		D_{voc}	D_{cons}	$D_{prévois}$	I	D_{asp}	T_1	F_o	Exemple
INITIALE	C V	-	-	+	+	+	+	+	bon / pont
INTERVOC	V C V	+	+	+	+	+	+	+	rebut / repu
FINALE	V C	+	+	+	+	-	-	-	bague / bac
PRECONS 1	V C $\underset{\smile}{C}$ V	+	+	-	+	-	-	-	bague / bac perdu(e)
PRECONS 2	V C $\underset{\smile}{C}$ V	+	+	-	+	-	-	+	bac / bague bizarre
PRECONS 3	V C$^{\neg}$ $\underset{\smile}{C}$ V	+	-	-	-	-	-	-	cab / cape perdu(e)
PRECONS 4	V C$^{\neg}$ $\underset{\smile}{C}$ V	+	-	-	-	-	-	+	cab / cape bizarre
INTERCONS 1	R C $\underset{\smile}{C}$ V	-	+	-	+	-	-	-	orgue / orque perdu(e)
INTERCONS 2	R C $\underset{\smile}{C}$ V	-	+	-	+	-	-	+	orgue / orque bizarre
INTERCONS 3	R C$^{\neg}$ $\underset{\smile}{C}$ V	-	-	-	-	-	-	-	serbe / serpe poli(e)
INTERCONS 4	R C$^{\neg}$ $\underset{\smile}{C}$ V	-	-	-	-	-	-	+	serbe / serpe bizarre

Tableau I - Matrice des allotraits du voisement

Les indices répertoriés sont ceux de la première occlusive C. Nous avons adopté les conventions suivantes : $\underset{\smile}{C}$: occlusive sourde $\overset{\frown}{C}$: occlusive sonore
$C^{\neg}$: occlusive non relâchée (sans détente)
R : son correspondant au phonème /R/

Comme indices neutralisés nous relevons :

- D_{voc} et D_{cons} à l'INITIALE
- D_{asp}, T_1, F_o à la FINALE
- $D_{prévois}$, D_{asp}, T_1 dans le cas d'une séquence V $\overset{\frown}{\underset{\smile}{k}}$ b V / V g b V (PRECONS 2)

où le [k] conserve sa détente et où l'assimilation de sonorité régressive (↶) doit permettre au Fo de la voyelle d'être différent selon qu'il s'agit d'un [k̥] ou d'un [g]. Pour V k p V / V g̊ p V ce n'est évidemment pas le cas, les variations de Fo sur la voyelle sont alors neutralisées (PRECONS 1). En outre, dans les contextes où il y a assimilation de détente (PRECONS 3 et 4) l'intensité I de l'explosion est évidemment neutralisée mais la question se pose de savoir si, le repère temporel servant à mesurer la durée de la tenue ayant disparu, l'indice D_{cons} est effectivement neutralisé; dans ce cas la durée totale du groupe consonantique pourrait être significative. Mais comme il a été constaté (DURAND, 1936; THORSEN, 1967; ROCHETTE, 1974) que les désonorisées s'allongeaient et les sonorisées s'abrégeaient (les quelques mesures disponibles ne semblent pas pouvoir ordonner autrement que sp, z̥p / zb, s̬b), nous considérons pour le moment l'indice D_{cons}, trop sensible à $D_{prévois}$, comme neutralisé dans ce cas. Dans la position PRECON 3 V b̥ p V, du fait de l'assourdissement du groupe, Fo ne fonctionne pas pour la dernière consonne, comme cela est encore possible pour PRECONS 4 V p̬ b V.

Dans les positions interconsonantiques avec [R] nous considérerons que lorsqu'il y a assourdissement et assimilation de détente, par exemple dans serp(e) perdu(e) / Serb(e) perdu l'opposition est véritablement relayée par le [R] : [sɛʀpp̥ɛʀdy / sɛʀb̥pɛʀdy] opposent [ʀ] voisé à [ʀ̥] dévoisé; les autres indices disparaissent et D_{voc} qui peut-être pourrait jouer un rôle semble devoir se traiter dans le cadre de [ʀ] / [ʀ̥] (INTERCONS 3). Quelques indices propres à la consonne en position INTERCONS réapparaissent lorsqu'il y a sonorisation : Fo pour INTERCONS 2 et 4, D_{cons} et I lorsqu'il n'y a pas assimilation de détente (INTERCONS 1 et 2).

Il est évident que certaines des données ci-dessus, qui nous sont parfois personnelles et quelque peu fragmentaires, auraient besoin d'être statistiquement mieux vérifiées, ne serait-ce qu'à l'analyse. Disons, pour être bref, que les besoins de la démonstration nous ont obligé à devancer quelque peu les données disponibles et à poser quelques hypothèses.

A l'examen global de la matrice, on s'aperçoit qu'il n'y a qu'une seule position où tous les indices sont présents : l'INTERVOCALIQUE; qu'il est possible qu'aucun n'apparaisse (INTERCONS 3); enfin, ce qui justifierait le caractère abstrait du trait : aucun indice n'est constamment présent.

Comment mener l'analyse de distinctivité d'une telle matrice ? Analogiquement il semble qu'il existe aussi des matrices *traits / allophones* qui ne présente aucune constante propre au phonème concerné. Ainsi, dans la matrice des allophones de /m/ en français (tableau II), le seul trait commun à toutes les positions *+ nasal* ne suffit pas à définir /m/.

		bilabial	*labiodental*	*occlusif*	*nasal*	*voisé*
	m	+	-	+	+	+
(-isme)	m̥	+	-	+	+	-
(femme volage)	ɱ	-	+	-	+	+
(femme fatale)	ɱ̥	-	+	-	+	-

Tableau II - Exemple de /m/ avec ses allophones

La solution réside évidemment dans la recherche d'une constante plus abstraite. On peut obtenir celle-ci dans l'analyse des incompatibilités en trouvant le trait commun à *bilabial* et *labiodental*, soit *labial*. Ainsi avec ses deux constantes : *nasal* et *labial* /m/ peut être défini par rapport à n'importe quel autre phonème.

Pourtant l'exemple classique des allophones de /R/ dans une variété de français régional (ABRY, 1977), sert à montrer qu'il peut exister un ensemble de réalisations d'une unité linguistique qui ne présente pas de corrélat constant.

	apical	*uvulaire*	*vibré*	*voisé*
r	+	-	+	+
r̥	+	-	+	-
ʀ	-	+	+	+
ʁ	-	+	-	+
ʁ̥	-	+	-	-

Tableau III - Exemple des allophones de /R/ pour une variété de français régional.

Cette fois-ci, il ne semble pas possible de réduire les deux colonnes incompatibles *apical* et *uvulaire* à un trait commun (le passage au niveau acoustique ne le permettrait pas davantage). L'invariant dans ce cas est habituellement cherché dans le comportement distributionnel de l'unité (complémentarité et variante libre). Avec l'abandon du critère de similarité phonétique, il n'y a guère qu'une expérimentation phonétique perceptive sur la substitution entre variantes, même lorsque celle-ci n'est pas le fait habituel du locuteur, qui puisse valablement nous tirer d'affaire dans ce cas.

Cette procédure peut-elle être transposée au niveau des indices ? Il est bien évident qu'il n'est pas toujours possible, même expérimentalement de substituer les ensemble d'indices les plus différents (c'est-à-dire les lignes les plus distantes de la matrice) alors qu'il était tout à fait possible d'introduire un [ɣ] à la place de sa variante complémentaire [r]. Par exemple comment pourrait-on réaliser des stimuli comportant les indices de l'INTERVOC à l'INTERCONS 3 ? On peut bien sûr rajouter l'explosion, mais pas T_1 évidemment; de même on ne peut rajouter D_{voc} à l'initiale. Par contre, il est possible d'arriver à ce résultat en opérant des substitutions successives entre ensembles relativement proches et en considérant la substitution des plus distants comme fondée transitivement.

La question de l'invariant évoquée, il reste une possibilité d'analyse de la redondance dans une matrice de ce type. Si celle de /R/ par exemple ne comporte pas de constantes, il n'en reste pas moins que la connaissance de certains traits permet d'en prévoir d'autres. L'analyse deux à deux des colonnes montre, outre l'incompatibilité déjà constatée, que + *apical* entraîne + *vibré*

et que tout - *vibré* est + *uvulaire*. Ce qui permet déjà de différencier ces deux traits qui ont une puissance de distinctivité plus grande que ceux qui n'impliquent rien, comme *voisé*.

On voit donc que ce type d'analyse que l'on peut faire lorsqu'il n'y a pas d'indice invariant, donne la possibilité de continuer à distinguer les traits d'après le critère de distinctivité puisqu'il apparaît des implications.

Nous proposons donc une analyse de la distinctivité sur la matrice des indices du voisement qui fasse apparaître les relations de présupposition. Voici les résultats sous forme de diagramme, les flèches indiquant les implications :

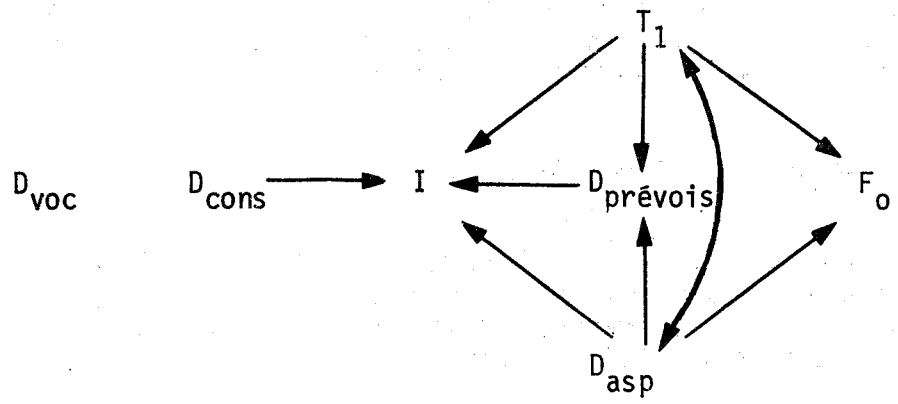


Tableau IV - Structure de redondance linguistique des indices.

En quantifiant chacun des noeuds par le nombre des flèches qui en partent, diminué du nombre de celles qui y arrivent, on obtient une échelle de distinctivité des indices (tableau V).

+3	T_1, D_{asp}
+2	
+1	D_{cons}
0	D_{voc}
-1	$D_{prévois}$
-2	F_o
-3	
-4	I

Comme actuellement il n'existe malheureusement pas de classement perceptif des indices comparable à la hiérarchie des traits établie, par exemple par MILLER & NICELY (1955), nous ne pouvons, pour toute comparaison, qu'utiliser l'exposé synthétique de DELATTRE (1965) pour l'anglais. Les quatre premiers indices T_1, D_{asp}, D_{cons} et D_{voc} correspondent pour DELATTRE à des choix perceptifs nets. Pour $D_{prévois}$ et I, ce n'est pas le cas (ce dernier indice ne semble guère avoir été expérimenté) (HOFFMAN, 1957). Quant à Fo, qui fait aussi partie des trois derniers, DELATTRE ne l'ayant pas utilisé, il faut bien reconnaître que les études qui lui ont été consacrées, n'ont pas établi de façon évidente son influence perceptive (CHISTOVITCH, 1969; HAGGARD & al., 1970; FUJIMURA, 1961; LARREUR & BOË, 1972).

Tableau V - Echelle de distinctivité des indices établie à partir des relations d'implication.

EN CONCLUSION ...

Il va de soi qu'une telle échelle, établie par la variance linguistique naturelle, n'est pas directement comparable à celle qui pourrait être obtenue par expérimentation psychophonétique. Un exemple tristement célèbre d'inadéquation flagrante, obtenue par analyse de distinctivité au niveau des traits, est l'attribution, par MARTINET, de la pertinence au trait *bilabial* et non au trait *occlusif* de /p/, le premier présupposant le second. L'interprétation que font les francophones des constrictives bilabiales [Φ] comme /f/ et non comme /p/ est un démenti à cette hypothèse. A notre avis une analyse détaillée des variantes phonétiques françaises aurait pu permettre d'éviter une telle erreur. Nous ne cachons pas non plus qu'en descendant l'analyse au niveau des indices, nous comptions bien diminuer ce risque. Il n'existe d'ailleurs pas d'autre alternative que d'abandonner la tâche au psychophonéticien qui de toute façon pourra utiliser ce type d'analyse :

1. pour montrer que tel indice ne saurait être l'indice prépondérant, tous contextes confondus;

2. pour formuler des hypothèses sur le traitement des indices : dépendant ou non du contexte (SERNICLAES, 1975) et élaborer un modèle de perception qui opère avec la notion d'invariant sous contexte, puisque l'examen des situations montre qu'aucun indice n'est partout présent;

3. ou bien pour orienter les recherches vers de nouveaux indices, invariants (DELBROCK, 1970) ou moins sensibles au contexte.

4. Enfin, généralement les hypothèses phonologiques clairement formulées doivent bien permettre de mettre en place les stratégies qui pourraient ou non les infirmer.

REFERENCES

ABRY, C., 1977, Distribution contextuelle et géographique de deux vibrantes phonologiques dans quelques parlers francoprovençaux; -Revue de Linguistique Romane 41, pp. 241-284.

CHISTOVICH, L.A., 1969, Variations of the fundamental voice pitch as a discriminatory cue for consonants; - Soviet Physics Acoustics 14, pp. 372-378.

DEBROCK, M., 1971, Les corrélats physique et physiologique de la force articulatoire des consonnes occlusives et constrictives initiales et intervocaliques; ITL 13, pp. 29-59.

DELATTRE, P., 1958, Les indices acoustiques de la parole. Premier rapport; - Phonetica 2, pp. 108-118 & 226-251.

DELATTRE, P., 1965, Comparing the phonetic features of English, French, German and Spanish; -Julius Groos Verlag, Heidelberg.

DELATTRE, P., 1967, Des indices acoustiques aux traits pertinents; -6th Int.Congr. Phonet. Sci., Prague, pp. 35-47.

FUJIMURA, O., 1961, Some synthesis experiments on stop consonants in initial position; -M.I.T. QPR 61, pp. 153 .

HAGGARD, M., AMBLER, S., & CALLOW, M., 1970, Pitch as a voicing cue; -J.A.S.A. 47, pp. 613-617.

HOFFMAN, H.S., 1957, A study of some cues in the perception of the voiced stop consonants; -Diss. Univ. Connecticut.

LARREUR, D., 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, pp. 103-126.

MILLER, G.A. & NICELY, P.E., 1955, An analysis of perceptual confusions among some English consonants; -J.A.S.A. 27, pp. 338-352.
trad. franç. in MEHLER, J. & NOIZET, G., 1974. Textes pour une psycholinguistique, pp. 175-207. Mouton. Paris - La Haye.

ROCHETTE, C.E., 1973, Les groupes de consonnes en français. Etude de l'enchaînement articulatoire à l'aide de la radiocinématographie et de l'oscillographie; pp. 371-372; 373-398, Klincksieck, Paris et les Presses de l'Université Laval, Québec.

SERNICLAÉS, W., 1975, Traitement indépendant ou interaction dans le processus d'intégration perceptive des indices du voisement ? ; 6° J.E. sur la Parole. Groupe Communication Parlée du GALF, Toulouse I, pp. 246-255.

THORSEN, O., 1967, Voice assimilation of stop consonants and fricative in French and its relation to sound duration and intra-oral air-pressure; A.R.I.P.U.C. I, pp. 67-76.

9èmes JOURNEES D'ETUDE SUR LA PAROLE

LANNION 31 mai - 2 juin 1978

ETUDE D'UN INDICE ACOUSTIQUE DES VOYELLES : LA PUISSANCE INTRINSEQUE

Bernard GUERIN — Laboratoire de la Communication parlée - E.N.S.E.R.G. 23 rue des Martyrs 38031 GRENOBLE Cedex

Louis-Jean BOË — Institut de Phonétique - Université des Langues et Lettres Domaine Universitaire B.P.25 Centre de Tri 38040 GRENOBLE Cedex

.RESUME

Toutes choses égales par ailleurs (contexte, accent, ...), l'intensité des voyelles présente en moyenne des différences statistiquement significatives.

Le but de ce travail est de montrer si l'origine de ces différences provient de la source vocale et/ou du conduit vocal. En utilisant des résultats déjà obtenus pour la fonction de transfert du conduit vocal, on a étudié sur le plan acoustique, et par simulation, le comportement du larynx et des cavités supraglottiques lors de la phonation de voyelles orales françaises.

On a montré que la puissance intrinsèque est liée à la fréquence du premier formant. Cette relation peut s'expliquer en tenant compte des caractéristiques du spectre de la source vocale (pente en -12 dB/octave) d'une part, de la fréquence du premier formant et de sa bande passante d'autre part.

STUDY OF A VOWEL ACOUSTIC FEATURE : INTRINSIC INTENSITY

Bernard GUERIN — Laboratoire de la Communication parlée - E.N.S.E.R.G. 23 rue des Martyrs 38031 GRENOBLE Cedex

Louis-Jean BOË — Institut de Phonétique - Université des Langues et Lettres Domaine Universitaire B.P.25 Centre de Tri 38040 GRENOBLE Cedex

SUMMARY

All other characteristics being equal (context, stress, ...), average vowel intensity shows statistically relevant value differences.

The purpose of this study is to show whether these differences are due to the vocal source or/and the vocal tract. Results relating to the vocal tract transfer function have been used to perform an acoustical study of the comportment of larynx and supraglottal cavities during french oral vowels phonation, by means of digital simulation.

It has been shown that the intrinsic intensity is related to the first formant, this relation can be explain by referring to vocal source spectrum characteristics (-12 dB/oct slope) on the one hand, and to both first formant and its bandwidth on the other hand.

ETUDE D'UN INDICE ACOUSTIQUE DES VOYELLES : LA PUISSANCE INTRINSEQUE

Bernard GUERIN Laboratoire de la Communication parlée - E.N.S.E.R.G.
23 rue des Martyrs 38031 GRENOBLE Cedex

Louis-Jean BOË Institut de Phonétique - Université des Langues et Lettres
Domaine Universitaire B.P.25 Centre de Tri 38040 GRENOBLE Cedex

1/ INTRODUCTION

Pour une même langue, dans des conditions contextuelles et prosodiques identiques, les niveaux d'intensité moyens des différentes voyelles sont significativement différents. Connue des phonéticiens depuis longtemps (ROUSSELOT , 1924) , cette constatation objective a été chiffrée pour l'anglais-américain par SACIA et BECK (1926), BLACK (1949, FAIRBANKS et al. (1950), LEHISTE et PETERSON (1959). Bien que l'on ne possède pas de résultats systématiques, les études menées, par exemple pour le français (MUNCH, 1969/70 ; ROSSI, 1971), le hongrois (FONAGY, 1966), le malgache (RAKOTOFIRINGA, 1968/69) et le polonais (WODARZ, 1961), donnent à penser qu'il s'agit d'un fait de phonétique générale. Les voyelles fermées sont moins intenses que les voyelles ouvertes ; les différences entre (i) et (ɑ) sont de l'ordre de 6 dB. Elles sont d'autant plus faibles entre deux voyelles que leur degré d'aperture est voisin.

Cette intensité, indice de timbre vocalique, est qualifiée d'intrinsèque dans la mesure où, étroitement liée au processus de production, elle n'intervient pas directement pour différencier les voyelles entre elles ; par ailleurs, elle est modifiée par l'environnement consonantique (HOUSE et FAIRBANKS, 1953 ; LEHISTE et PETERSON, 1959 ; RAKOTOFIRINGA, 1968) et par les faits suprasegmentaux, c'est à dire l'accent et l'intonation.

Cependant, l'existence de cet indice pose un certain nombre de problèmes que nous évoquerons brièvement :

* Il faut en tenir compte au cours de l'analyse instrumentale de l'accent ; ainsi, un (i) accentué peut être moins intense qu'un (ɑ) non accentué duquel il se différencie perceptivement. Ce problème a été largement abordé par LEHISTE et PETERSON (1959), WODARZ (1961), ROSSI (1967) et RAKOTOFIRINGA (1969).

* Puisque deux voyelles telles que (i) et (ɑ) sont perçues isophones, tout en ayant des intensités objectives différentes, on peut se demander si la perception de l'intensité se fait, dans ce cas, directement à partir du signal acoustique (LANE et al., 1961), ce que tendraient à prouver de judicieuses évaluations de leur phonie (ROSSI, 1971) ou à partir de la restitution par l'auditeur de l'effort articulatoire du locuteur, identique dans les deux cas (LEHISTE et PETERSON, 1959 ; LADEFOGED et Mc KINNEY, 1963).

On voit donc que l'étude de cet indice présente un certain intérêt ; l'objet de ce travail se situe en amont de ces problèmes d'analyse prosodique ou de mise à l'épreuve de tel ou tel modèle de perception. Il s'agit pour nous d'étudier la cause physiologique de ces différences d'intensité au niveau de la production. En effet, si toutes les mesures ont permis de mettre en évidence que les voyelles fermées sont moins intenses que les voyelles ouvertes, il n'existe pas, à notre connaissance, d'explications globales concernant ce phénomène.

En 1950, FAIRBANKS a essayé de dégager des relations entre la forme du conduit vocal et les caractéristiques intrinsèques d'intensité. Sur des coupes sagitales, schématisées à partir de radiographies, ont été mesurées les distances entre : . les deux lèvres,
. les incisives,
. le point le plus élevé de la langue et le palais (c'est à dire ce que les phonéticiens appellent l'aperture au niveau du lieu d'articulation).
Les statistiques révèlent que l'intensité des voyelles est bien corrélée avec la deuxième dimension, un peu moins bien avec la première et pas du tout avec la troisième. En fait, et on ne peut le reprocher à son auteur, cette démarche relève à la fois d'une méthodologie limitée et de la conception d'alors concernant la production de la parole. En effet, les coupes sagitales du conduit vocal ne permettent que la mesure d'une dimension, alors que l'on sait que c'est l'aire qui est importante et, d'autre part, aucune partie du conduit vocal n'est à elle seule responsable de son comportement acoustique global.

Nous avons étudié l'intensité intrinsèque des voyelles à partir des résultats de certaines études physiologiques et nous avons effectué une simulation de l'ensemble source - conduit vocal. Pour la source, nous avons choisi le modèle à deux masses d'ISHIZAKA et FLANAGAN (1972) que nous avons spécialement étudié par ailleurs (GUERIN et BOË, 1977 ; BOË et GUERIN, 1977). Rappelons simplement que le modèle est commandé par deux paramètres : la pression sub-glottique P_S et un paramètre Q qui rend compte à la fois de la tension et de l'amincissement des cordes vocales. En ce qui concerne le conduit vocal couplé à la source, nous nous sommes servis de tous les résultats de MRAYATI (1976) qui a plus spécialement étudié le problème des pertes ainsi que l'évaluation de la fonction d'aire pour les voyelles du français.

2/ SIMULATION NUMERIQUE DE LA PRODUCTION DES VOYELLES ORALES - EVALUATION DE LA PUISSANCE INTRINSEQUE

Par nos études antérieures, nous disposons d'une simulation numérique de la source vocale à deux masses couplée au conduit vocal (GUERIN et BOË, 1977). Le conduit vocal est représenté par son impédance d'entrée. MRAYATI (1976) a montré que cette impédance est équivalente à des circuits R.L.C. parallèles accordés sur les fréquences de formants, les pertes du conduit vocal étant représentées par l'amortissement de ces circuits. Cette simulation permet donc d'évaluer l'onde de débit de la source vocale $U_g(t)$ dans les conditions normales de phonation. A l'aide d'un programme de transformée de Fourier rapide, on peut obtenir le spectre de ce signal $U_g(t)$. Par ailleurs, on dispose de programmes permettant de calculer la fonction de transfert du conduit vocal : $H(\omega) = P(\omega)/U_g(\omega)$, où $P(\omega)$ est la transformée de Fourier de l'onde de pression aux lèvres et $U_g(\omega)$ la transformée de Fourier de l'onde de débit de la source vocale. Ces programmes tiennent compte de toutes les pertes dans le conduit vocal (MRAYATI, 1976).

Le puissance de l'onde de pression aux lèvres peut s'écrire :

$$P_u = \frac{1}{T} \int_0^T p^2(t)\, dt \quad .$$

Or p(t), pour une voyelle, est un signal périodique, on peut donc en prendre la transformée de Fourier, soit :

$$P(n\Omega) = \frac{1}{T_0} \int_0^{T_0} p(t)\, e^{-j\frac{2\pi nt}{T_0}}\, dt \quad (\text{avec } \frac{n}{T_0} = \omega),$$

et $T_0 = 1/\Omega$ est la période de base de p(t).

On peut alors écrire (formule de Parseval) :

$$P_u = \frac{1}{T} \int_0^T p^2(t)\, dt = \sum_{n=1}^{\infty} P^2(n\Omega) .$$

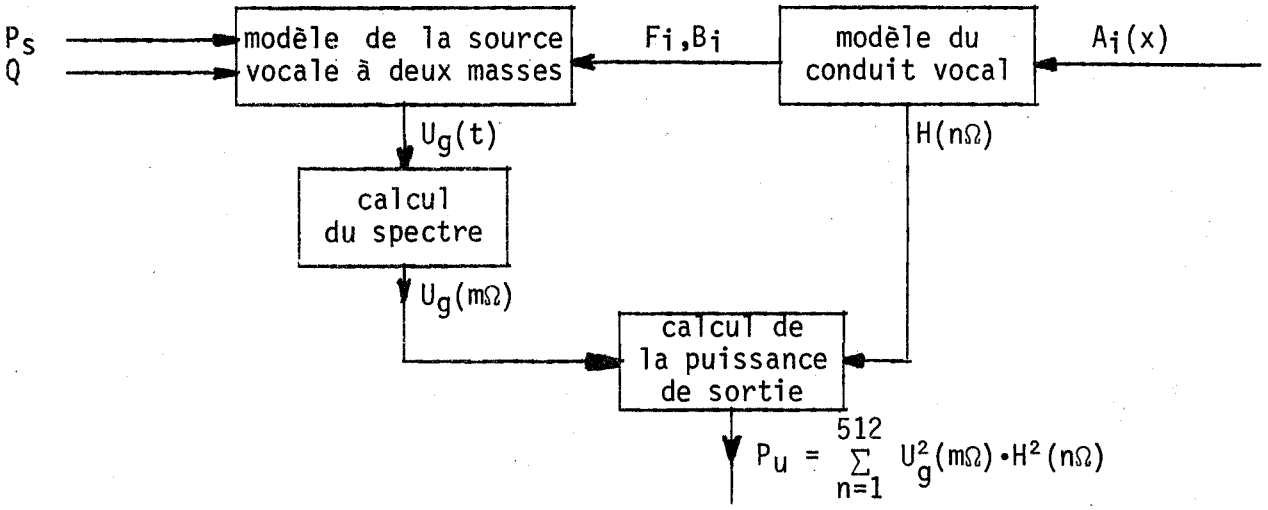
D'autre part, nous avons : $P^2(n\Omega) = H^2(n\Omega) \cdot U_g^2(n\Omega)$

donc : $P_u = \sum_{n=1}^{\infty} H^2(n\Omega) \cdot U_g^2(n\Omega)$

où $H(n\Omega)$ sont les composantes aux fréquences $n\Omega$ de la fonction de transfert du conduit vocal ; $U_g(n\Omega)$ sont les composantes harmoniques de l'onde de débit de la source vocale.

Connaissant les spectres de l'onde de débit $U_g(t)$ pour la source chargée de l'impédance d'entrée du conduit vocal d'une voyelle, et la fonction de transfert de cette voyelle, on pourra calculer la puissance P_u de la pression aux lèvres (figure 1).

FIGURE 1. Organigramme du calcul de la puissance de l'onde de sortie aux lèvres pour des voyelles orales françaises.
Computation flow chart of vocal tract output wave intensity for french oral vowels.



Pratiquement, nous avons limité le domaine des fréquences à 5 kHz, ce qui correspond à un échantillonnage temporel du signal de 10 kHz. Des expériences ont montré que l'énergie de p(t) au-delà de 5 kHz est toujours négligeable pour les voyelles.

Dans la suite, nous donnerons aussi la valeur de la puissance obtenue en considérant la source vocale comme non couplée au conduit vocal et dont le spectre présente une chute de 12 dB/octave au-delà de 100 Hz.

3/ ETUDE DU PHENOMENE

Dans le cas de la production des voyelles, le signal de parole est le résultat d'un seul signal source (l'onde de débit glottique) modifié par les cavités supraglottiques, c'est à dire en termes acoustiques, par sa fonction de transfert. En toute bonne logique, les caractéristiques intrinsèques

d'intensité peuvent donc être attribuées à plusieurs causes qui, d'ailleurs, ne s'excluent pas mutuellement. Nous avons considéré deux hypothèses :

PREMIERE HYPOTHESE :
Les caractéristiques intrinsèques d'intensité sont imputables à la source.

Le comportement de la source, c'est à dire l'intensité et/ou le spectre du signal glottique, est différent selon la nature de la voyelle. Il s'agirait, dans ce cas, d'un phénomène de couplage source - conduit vocal qu'il faudrait expliquer en termes acoustiques (ou physiologiques).

A/ Selon la disposition articulatoire, l'intensité du signal source est différente.

L'intensité du signal source est essentiellement due à la pression sub-glottique, la tension des cordes vocales ne modifiant que leur régime de vibration. Les mesures de LADEFOGED et Mc KINNEY (1963) et LADEFOGED (1967) montrent que, dans les mêmes conditions contextuelles et prosodiques, les locuteurs génèrent la même pression sub-glottique, quelle que soit la nature de la voyelle. Par contre, à même P_s, l'intensité du signal de parole varie avec des timbres vocaliques différents.

Au niveau de la simulation, nous avons retrouvé ce résultat. L'intensité du signal de la source a été mesuré pour deux ensembles de commande et pour des couplages correspondant à toutes les voyelles orales du français. Si nous prenons comme référence la voyelle la plus intense, c'est à dire la voyelle postérieure (ɑ), ce que nous ferons d'ailleurs systématiquement par la suite, on constate de très faibles différences entre les niveaux d'intensité (N.I.) du signal source (tableau 1).

TABLEAU 1. Niveau d'intensité du signal source, valeurs référencées par rapport à (ɑ), pour deux ensembles de commandes : (1) : P_s = 8 cm d'H_2O et Q = 2 ; (2) : P_s = 6 cm d'H_2O et Q = 1,5.

	N.I. du signal source en dB										
	u	o	ɔ	ɑ	a	ɛ	e	i	y	ø	œ
(1)	0	0,3	0	0	0	0,1	0,4	-0,1	0	0,4	0,1
(2)	0,5	0,2	0,1	0	0	0,1	0,2	0,2	0,5	0,2	0,1

Cette hypothèse semble donc pouvoir être rejetée : les différences d'intensité vocaliques ne se retrouvent pas au niveau de l'intensité de la source.

B/ Le spectre du signal source, caractérisé par sa pente, est différent selon la disposition articulatoire.

Si tel était le cas, des variations de la pente du spectre pourraient provoquer des différences d'intensités du signal de parole, même si les fonctions de transfert des voyelles n'étaient pas différenciées.

A notre connaissance, il n'existe pas beaucoup de travaux effectués sur de la parole naturelle permettant d'avoir une idée précise. Seules, les mesures de TAKASUGI (1971) laissent entrevoir qu'il y a peu de différences entre (i), (ɑ) et (u).

Par simulation, nous avons quand même observé une différence de 1 dB/oct. entre les pentes de (i) et (u) d'une part et (ɑ) d'autre part. Ce phénomène est dû au couplage entre le larynx et le conduit vocal ; il dépend essentiellement de la valeur du premier formant et se manifeste par de légères modifications de la forme de l'onde glottique (MRAYATI, 1976 ; GUERIN et al., 1976 ; GUERIN et BOË, 1977).

L'hypothèse de l'influence du couplage, modifiant la pente du signal glottique et provoquant un résultat quelque peu différent à la sortie du conduit vocal, n'est pas totalement à écarter et nous y reviendrons.

DEUXIEME HYPOTHESE :
C'est au niveau de la fonction de transfert du conduit vocal qu'il faut chercher la cause des indices d'intensité vocalique en tenant compte de la valeur des formants et de leurs bandes passantes, le tout par rapport à la répartition énergétique du signal glottique.

La source possède un spectre dont la pente décroît d'environ 10 dB/oct. Ce phénomène, s'il agissait seul, aurait tendance à favoriser l'intensité des voyelles à premier et deuxième formants bas. Mais il faut tenir compte de l'influence des pertes qui, par contre, sont d'autant plus importantes que le premier formant a une fréquence peu élevée.

En fait, il y a dans le conduit vocal quatre types de pertes : par viscosité, chaleur, vibration des parois et rayonnement aux lèvres. Pour des valeurs de la fréquence correspondant au premier formant (plus précisément inférieures à 500 Hz), ce sont les vibrations des parois qui apportent la plus grosse contribution à la valeur de la bande passante. Celle-ci augmente très vite si F_1 diminue (MRAYATI, 1976). Aussi, il faut s'attendre à une intensité intrinsèque fonction de F_1.

Avant d'utiliser les résultats de simulation, nous avons repris ceux de LEHISTE et PETERSON pour l'anglais - américain et de ROSSI pour le français.

TABLEAU 2. Caractéristiques intrinsèques des voyelles de l'anglais-américain, en dB par rapport à (ɑ) (LEHISTE et PETERSON)

	i	I	ɛ	æ	ɑ	ɔ	ʌ	U	u	ɜ
F_1	270	390	530	660	730	570	640	440	300	490
N.I.	-5,6	-4,4	-2,3	-2,7	0	-0,1	-2,2	-2,4	-5,3	-4,3

TABLEAU 3. Caractéristiques intrinsèques des voyelles du français, en dB par rapport à (ɑ) (ROSSI)

	i	e	y	u	ɛ	o	œ	ɔ	ɑ
F_1	240	360	240	280	640	420	480	600	680
N.I.	-4,5	-4,0	-2,5	-2,3	-1,7	-1,7	-1,2	-1,0	0

Les coefficients de corrélation entre F_1 et $1/F_1$ et les intensités intrinsèques des voyelles sont significatifs au seuil 0,01. La deuxième hypothèse semble donc se confirmer. A titre d'illustration, nous présentons les figures 2 et 3 et donnons les résultats des régressions linéaires pour ces deux ensembles vocaliques :

$N.I._{dB} = 10{,}6.10^{-3}\ F_1 - 8{,}27$ pour les voyelles de l'anglais-américain

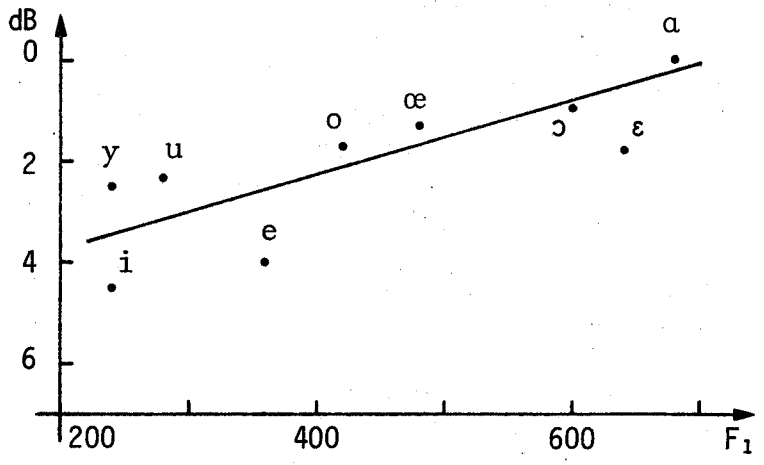
$N.I._{dB} = 6{,}4.10^{-3}\ F_1 - 4{,}93$ pour les voyelles du français

voyelle de référence : (ɑ).

Evolution de l'intensité des voyelles orales françaises en fonction de la fréquence du premier formant (d'après ROSSI, 1967)

FIGURE 2.

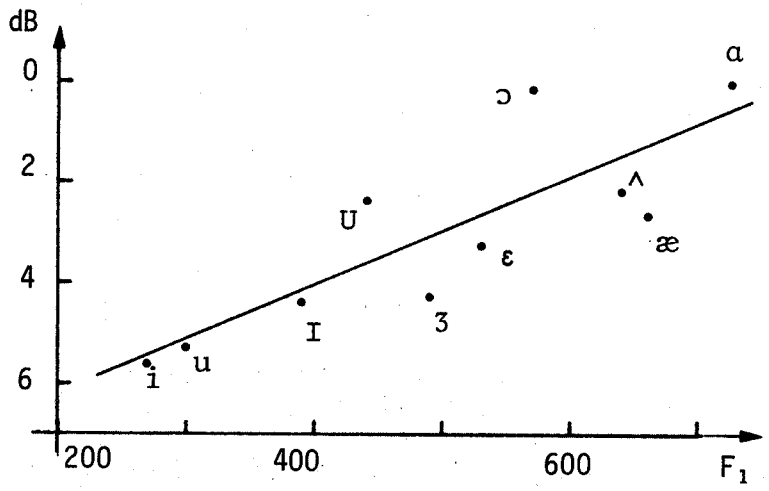
Variation of intensity of french oral vowels in relation with the first formant frequency (after ROSSI, 1967).



Evolution de l'intensité des voyelles orales anglo-américaines en fonction de la fréquence du premier formant (d'après LEHISTE et PETERSON, 1959)

FIGURE 3.

Variation of intensity of anglo-american oral vowels in relation with the first formant frequency (after LEHISTE and PETERSON, 1959).



Pour obtenir une confirmation des résultats obtenus sur la parole naturelle, nous avons opéré des simulations de comportement du conduit vocal en mesurant l'intensité du signal émis, pour les configurations vocaliques du français, avec une source fixe ayant un spectre à -12 dB/octave, et avec le modèle à deux masses, dont on sait par ailleurs qu'il simule le couplage source-conduit vocal.

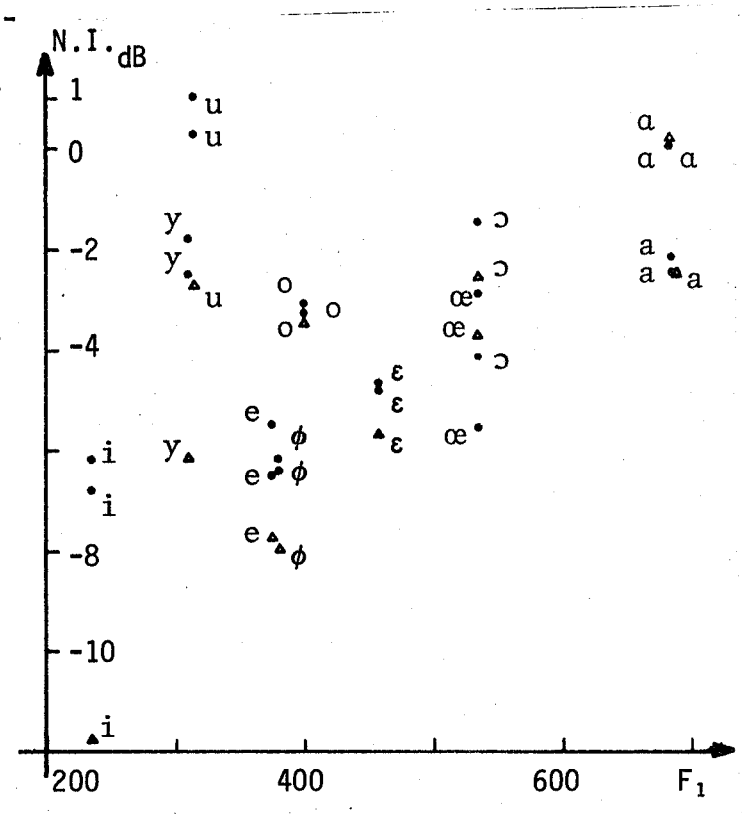
TABLEAU 4. Simulations pour les voyelles du français, les niveaux d'intensité ayant été chiffrés par rapport à la voyelle de référence (ɑ). Le conduit vocal a été excité par une source à spectre constant (-12 dB/octave) et par le modèle à deux masses qui permet de rendre compte des effets de couplage dans deux conditions :
(1) $P_S = 8$ cm d'H_2O, $Q = 2$; (2) $P_S = 6$ cm d'H_2O, $Q = 1,5$.

	u	o	ɔ	ɑ	a	ɛ	e	i	y	ø	œ
F_1 (Hz)	265	400	539	676	677	458	378	238	260	382	534
B_1 (Hz)	75	43	38	36	35	34	37	72	69	48	35
N.I. sans couplage (dB)	- 2,7	- 3,5	- 2,6	0	- 2,5	- 5,7	- 7,8	-11,8	- 6,1	- 8,0	- 3,7
avec couplage (1)	0,3	- 3,1	- 4,2	0	- 2,2	- 4,9	- 6,5	- 6,2	- 2,7	- 6,3	- 5,6
avec couplage (2)	1,0	- 3,4	- 1,4	0	- 2,7	- 4,9	- 5,5	- 6,9	- 1,9	- 6,2	- 2,8

Evolution de l'intensité des voyelles orales françaises en fonction de la fréquence du 1er formant et dans les 3 conditions de simulation définies au tableau 4

FIGURE 4.

Variation of french oral vowels intensity in relation with the 1st formant frequency and under the 3 simulation conditions defined in table 4.



De ces résultats, nous pouvons tirer les observations suivantes (fig.4) :

☆ En valeur absolue et en évolution, nous retrouvons bien les valeurs des caractéristiques intrinsèques. Une exception, le (u), perturbe quelque peu cette confirmation ; nous réservons son cas pour le moment : il nous obligera à mener une minutieuse vérification de la fonction d'aire et des pertes (il faudrait également

tenir compte du second formant).
☆ Pour les trois simulations, les intensités intrinsèques sont corrélées deux à deux significativement au seuil 0,01 : le couplage n'apporte pas de modifications importantes par rapport à l'action globale du conduit vocal.
☆ Sans couplage, les intensités intrinsèques sont, comme pour la parole naturelle, corrélées (au seuil 0,01) avec F_1 et $1/F_1$. Comme nous disposons des valeurs des bandes passantes, il est intéressant de pouvoir les relier au phénomène étudié : on note des corrélations significatives entre les intensités de chaque voyelle et F_1/B_1 ou $1/(F_1.B_1)$.
☆ Le couplage modifie un peu quantitativement ces évaluations : on ne les retrouve significatives qu'au seuil 0,05. Sans pour cela remettre en cause notre démonstration, il semblerait que le couplage soit dans la simulation surestimé pra rapport à son action dans la parole naturelle.

4/ CONCLUSIONS

Nos diverses constatations, tant sur la parole naturelle que simulée, nous conduisent à penser que le conduit vocal est à l'origine des caractéristiques intrinsèques des voyelles. Compte tenu du spectre de la source (-12 dB/octave), les pertes relatives au premier formant ont une influence prépondérante et, parmi celles-ci, les vibrations des parois. Les constatations de FAIRBANKS peuvent être interprétées pour des voyelles dont le premier formant (dont nous avons vu l'importance prépondérante) est sensible aux aires des sections du conduit vocal situées au-delà de la constriction : il suffit de consulter les fonctions de sensibilité du conduit vocal (MRAYATI, 1976).

On peut donc dire qu'il n'y a pas de relations directe avec les dimensions de l'une des constrictions mais que ces dernières interviennent de par leur influence sur le premier formant.

BIBLIOGRAPHIE

BOË L.J., GUERIN B. (1977) I.P.S.77 Miami
FAIRBANKS G. (1950) Speech Monogr. 17, 390-395
FONAGY I. (1966) J.S.H.R. 9, 231-244
GUERIN B., MRAYATI M., CARRE R. (1976) I.E.E.E. I.C.A.S.S.P., 47-50
GUERIN B., BOË L.J. (1977) 9ème I.C.A., I.76
HOUSE A.S., FAIRBANKS G. (1953) J.A.S.A. 25, 105-113
ISHIZAKA K., FLANAGAN J.L. (1972) B.S.T.J. 51, 1233-1268
LADEFOGED P., Mc KINNEY N.P. (1963) J.A.S.A. 35, 454-460
LADEFOGED P. (1967)
Three areas of experimental phonetics, Oxford Univ. Press, London
LANE H.L., CATANIA A.C., STEVENS S.S. (1961) J.A.S.A. 33, 160-167
LEHISTE I., PETERSON G.E. (1959) J.A.S.A. 31, 428-435
MRAYATI M. (1976) Thèse d'Etat, Université Scient. et Médic. Grenoble
MUNCH J. (1969-70) Travaux de l'Inst. de Phonét. de Strasbourg, 2, 112-125
RAKOTOFIRINGA H. (1967) Bulletin de l'Inst. de Phonét. et de Linguist. Grenoble
ROSSI M. (1967) Proc. 6th Int. Congr. Phonetic Sciences, 779-786
ROSSI M. (1971) Phonetica, 24, 129-161
ROUSSELOT P.J. (1924)
Principes de phonétique expérimentale, H. Didier, Paris
SACIA C.F., BECK C.J. (1926) B.S.T.J. 5, 393-403
TAKASUGI T. (1971) Jl of the Radio Research Laboratories, 18, 97, 209-220
WODARZ W. (1961) Phonetica 6, 177-215.

° °
°

9èmes JOURNEES D'ETUDE SUR LA PAROLE

LANNION 31 mai - 2 juin 1978

LABIALITE VOCALIQUE ET LABIALITE CONSONANTIQUE EN FRANCAIS

Premiers résultats

R. DESCOUT — CNET - Lannion
L.J. BOË - C. ABRY — Institut de Phonétique - Grenoble

.RESUME

La labialité en français est un trait de mode à la fois pour les voyelles (y, ø .../i, e...) et les consonnes (ʃ, ʒ.../s, z...). Nous avons étudié le comportement de ce trait dans le cas où, du fait des règles de coarticulation, la labialité des consonnes qui est non phonologique assimile les voyelles phonologiquement labiales.

Le corpus est constitué de mots où figurent en position finale accentuée les syllabes CV (s, z, ʃ, ʒ . * i, e, y, ø), soit 96 réalisations pour un locuteur. Les mots ont été placés dans des phrases porteuses.

Un labiofilm, face et profil, (35 mm, 50 images/sec., son synchrone) a été réalisé pour 5 locuteurs (2 femmes, 3 hommes).

Les paramètres retenus sont : l'écartement aux commissures (A), l'aperture entre les lèvres (B), l'aire A.B (C), l'aperture à la sortie du conduit vocal (D), la protrusion des lèvres (F), supérieure (F_1), inférieure (F_2)

Les mesures acquises numériquement (agrandisseur, tablette d'entrée graphique) ont été traitées statistiquement.

Les résultats montrent que : - les paramètres les plus significatifs sont, pour les voyelles comme pour les consonnes A et B, bien davantage que la protrusion F.

- deux paramètres sont nécessaires pour différencier les consonnes. A et B étant en outre significativement corrélés - et ceci seulement pour les consonnes - la présence d'un facteur de forme différencie la labialité consonantique de la labialité vocalique.

- pour les voyelles par contre un seul paramètre suffit : B reflète directement la distinction phonologique + lab/ - lab, A permet en plus de différencier pour les voyelles [- lab] 2 degrés de labialisation

- l' effet de coarticulation consonne-voyelle montre une assimilation progressive pour A, regressive pour B.

VOWEL AND CONSONANT LIP-ROUNDING IN FRENCH - First Results -

R. Descout — CNET - Lannion
L.J. Boë - C. Abry — Institut de Phonétique - Grenoble

. SUMMARY

Rounding in French is a manner feature for vowels (y, ø, ... vs. i, e ..) as well as for consonants (ʃ, ʒ vs. s, z). We study in this paper lip-rounding in the particular context where, because of coarticulation rules, non phonological consonant rounding influences phonological vowel rounding.

Our corpus consists of 96 short sentences containing in their last syllables CV [± lab] combinations : [s, z, ʃ, ʒ] with [i, e, y, ø].

A labiographic film (front and profile, 50 frames/sec., 35 mm, with sound synchronization) has been shot for 5 subjects.

Parameters measured are :
A, width of lip opening,
B, height of lip opening (front),
C, area A.B of lip opening,
D, distance between outermost points of lips,
F, protrusion of upper (F_1) and lower (F_2) lip.

Data were digitalized and statistically processed :

- for vowels as well as for consonants, the most significant parameters are A and B (and not protrusion F).
- for consonant discrimination, two parameters are needed. More, A and B showing significant correlation - and only for consonants - we can say that consonant rounding differs from the vowel one by the presence vs. absence of a specific form factor.
- on the contrary, for vowel discrimination, we need only one parameter :

B. reflects directly the phonological distinction *+ lab / - lab*
A. can operate a more precise discrimination, but for [- lab] vowels only. We can so distinguish between two degrees of lip-rounding.

- CV coarticulation effects show that rounding is left-to-right for A and right-to-left for B.

LABIALITE VOCALIQUE ET LABIALITE CONSONANTIQUE EN FRANCAIS

Premiers résultats

R. DESCOUT — CNET - Lannion
L.J. BOË - C. ABRY — Institut de Phonétique - Grenoble

INTRODUCTION

Parmi les traits articulatoires qui relèvent à la fois des consonnes et des voyelles, la labialité occupe en français une place importante.

Au niveau phonologique, elle participe en tant que trait de lieu à la définition des consonnes occlusives et constrictives /p, b, m, f, v/, et comme trait de mode à celle des voyelles /y, ø, œ, œ̃/. Mais ce n'est qu'au niveau phonétique que cette modalité labiale fonctionne également pour les consonnes. On retrouve, en effet, le même type de mimique labiale pour les deux constrictives [ʃ, ʒ], mais ce trait reste phonologiquement redondant. Et pourtant au niveau des règles de coarticulation c'est bien ce même trait redondant que les chuintantes imposent dans une assimilation progressive : ainsi le /i/ de /ʃi/ sera-t-il labialisé : soit au niveau phonétique [ʃ͡ʲi]. Ce comportement est d'autant plus remarquable que dans les séquences CV l'assimilation en français est habituellement régressive (cf Tableau I) : ce sera le cas par exemple pour /sy/ réalisé [s͡ʲy] .

niveau	s	i	s	y	ʃ	i	ʃ	y
phonologique	o	-	o	+	o	-	o	+
phonétique segmental	o	-	o	+	+	-	+	+
phonétique combinatoire	-	-	+	+	+	+	+	+

Trait de labialité : + : labial
- : non labial
o : non défini

TABLEAU I

Au niveau phonétique chaque son est considéré avant l'application des règles de coarticulation (segmental) et après celles-ci (combinatoire)

A première vue, une telle assimilation semble poser problème quant à la préservation de la distinction de labialité entre voyelles précédées de ʃ et ʒ ; et pourtant, même dans ce contexte, l'opposition *voyelle labiale / voyelle non labiale*, ne fait aucun doute pour un francophone : /ʃi/ ~ /ʃy/

réalisé [ʃi̯] ~ [ʃy] maintient la différence entre /i/ et /y/.

Sans qu'il soit nécessaire d'envisager dès maintenant les variations concomitantes de lieu et d'aperture qui opposent le /i/ au /y/, conjointement au trait de labialité, cette contradiction entre sa redondance et sa pertinence a de fortes chances de n'être qu'apparente :

. En effet le mode de raisonnement phonologique dont elle provient ne fait pas de degré intermédiaire entre pertinence et redondance : au niveau perceptif, il n'est pas dit que la protrusion labiale pour [ʃ, ʒ] ne puisse pas intervenir à côté des indices fournis par le lieu d'articulation (ou plus exactement par la forme de la constriction).

. De même en ce qui concerne le trait de labialité, l'habitude de le traiter phonologiquement comme une opposition privative binaire (*+ lab./ - lab.*) élimine la possibilité d'utiliser des distinctions tenant à différents degrés de labialité. Ne peut-on avoir par exemple, plus de deux classes pour les voyelles i et y ? Dans ce cas le trait serait plutôt scalaire que binaire :

[si]	non labialisé	0
[ʃi̯]	moyennement labialisé	1
[ʃy et ʃy̯]	très labialisé	2

. Enfin il est possible de penser qu'il existe deux natures articulatoires différentes pour la labialité suivant qu'elle se manifeste comme trait de mode dans les consonnes ou dans les voyelles.

Le but de cette étude sera ainsi de spécifier la nature de la labialité consonantique et vocalique dans le cas particulier où, mises en coarticulation, elles apparaissent phonologiquement en contradiction.

METHODE

Corpus

Nous avons limité notre corpus, dans un premier temps, aux syllabes ouvertes finales, qui sont normalement accentuées. Dans le but d'obtenir, pour comparaison, les quatre combinaisons CV possibles entre les deux valeurs du trait de labialité (cf. Tableau I), le choix de la consonne [s] articulatoirement la plus proche de [ʃ] s'imposait. Un possible effet du voisement nous a fait introduire [z] et [ʒ]. Les oppositions de labialité vocalique retenues sont i / y et e / ø (œ n'existant pas en finale absolue et la combinaison [chuintante + œ̃] n'étant pas attestée en français, sauf ʒœ̃). Le corpus a été ainsi constitué de paires du type *branchie / branchue, concis / conçu* pour i/y, *rassis / rachis, déçu / déchu* pour s/ʃ et de quasi-paires intermédiaires type *machie / massue, déci / déchu*, soit 6 oppositions pour i, y et 6 pour e, ø ; ceci pour les sourdes et autant avec les sonores z,ʒ. Notre propos n'étant pas, pour l'essentiel, de spécifier les oppositions d'aperture nous n'avons pas introduit de paires spécialement pour i/e et y/ø. Chaque mot a été placé dans une phrase porteuse du type : *C'est bien concis/ conçu*. Les 48 paires ont été disposées en ordre aléatoire, ce qui nous fournit 96 réalisations CV.

Sujets

Cinq locuteurs francophones non marqués : 3 hommes et 2 femmes sans antécédents pathologiques de la face et du conduit vocal autres que de banales inter-

ventions dentaires, ont été choisis.

Appareillage

Les cinq sujets ont été maquillés (lèvres blanchies, dents noircies) avec des traits et points de repère cutanés (plan de Francfort, 4 points aux lèvres et 1 au menton) pour la prise de vue simultanée de face et de profil avec un miroir à 45° (film 35 mm noir et blanc à 50 images/sec.*; son synchronisé sur magnétophone grâce à un "clap optique" repérable aussi sur la piste son et inséré entre chaque phrase). Les sujets avaient leur tête appuyée et maintenue immobile.

Dépouillement et mesures

Les images repérées (par rapport au clap optique) correspondant au centre des réalisations consonantiques et vocaliques ont été projetées avec un agrandisseur sur une tablette d'entrée graphique reliée à un ordinateur. Les contours dessinés manuellement sont acquis numériquement** et mémorisés. Les paramètres retenus sont : (figure 1)

-de face : A l'écartement aux commissures.

B l'aperture intérolabiale.

C l'aire.

-de profil : D l'aperture extérolabiale à l'extrémité du conduit vocal.

F_1 et F_2 mesurent respectivement la protrusion-rétraction des lèvres supérieure et inférieure par rapport à la position préphonatoire.

Certains paramètres peuvent sembler redondants (D et B, F_1 et F_2); mais outre qu'il est nécessaire de quantifier empiriquement leur éventuelle corrélation en ce qui concerne la labialité de mode (avec l'intervention de la labialité de trait pour [fy] par exemple, cette corrélation ne serait plus aussi évidente, la rétraction de la lèvre inférieure pour la consonne interférant avec la protrusion pour la voyelle), certains sont nécessaires pour le calcul des aires aux lèvres permettant de passer aux mesures acoustiques. Il est aussi évident, par exemple, qu'un facteur de forme différencie la labialité consonantique "en pavillon" et l'oppose aux formes "en amande" plus ou moins arrondies des voyelles : cette forme ne peut donc être suffisamment approchée à partir des seuls A et B.

RESULTATS ET DISCUSSION

Les résultats ci-dessous sont ceux dont nous disposons actuellement pour un seul locuteur (sujet féminin).

Labialité consonantique

La figure 2 présente les tracés moyens pour [ʃ,ʒ] et [s,z] suivant les contextes vocaliques.

---

*. Le film a été réalisé par l'équipe Cinéma du S.I.R.P. (Service d'Information et de Relations Publiques des PTT).

** Nous avons utilisé un programme mis au point pour ce type d'acquisition par S. MAEDA.

I. Discrimination *sourde / sonore* . Si on considère [ʃ, ʒ] d'une part et [s, z] d'autre part, tous contextes vocaliques confondus, le voisement n'apporte pas de différence significative.

Il est à noter que ceci semble aller contre de très anciens résultats puisque les deux photographies reprises du Traité de prononciation française de BURGUET, que ROUSSELOT publiait dans ses Principes ... (t.I, p. 588), montrent une nette différence au moins d'aperture, ʃ étant plus ouvert que ʒ.

2. Discrimination ʃ, ʒ / s, z. Aucun paramètre à lui seul ne permet de faire la distinction* entre [ʃ, ʒ] d'une part et [s , z] de l'autre, tous contextes vocaliques confondus. On pourrait s'étonner de ce que la protrusion en particulier ne puisse pas opérer cette discrimination (figure 3) si l'on ne se rappelait que [s$_{\omega}$] et [z$_{\omega}$] peuvent être aussi protrus.

Par contre, avec deux paramètres il est possible de faire la partition* :

- avec A et B soit l'écartement et l'aperture,
- avec F_1 ou F_2 et B,
- avec F_1 ou F_2 et D,

la meilleure de toutes ces partitions paraissant être A et B (figure 4). Par contre :

- F_1 (ou F_2) et A,
- A et D,
- B et D,

ne le permettent pas.

3. Relations entre les paramètres. Il existe un coefficient de corrélation significatif (seuil 0,01) pour :

- F_1 et F_2, ce qui n'est pas surprenant lorsqu'on sait que dans les conditions de notre corpus, il n'y a pas de labialité de trait qui pourrait interférer pour la lèvre inférieure (cas de [f, v])et/ou supérieure [p, b, m] avec la labialité de mode;

- B et D, l'aperture entre les lèvres étant elle aussi évidemment corrélée à l'aperture en bout du conduit vocal.

Plus intéressantes paraissent être les corrélations suivantes :

- F_1 (ou F_2) et B, montrant que lorsque l'aperture croit la protrusion décroit (figure 3),

- et surtout A et B, corrélation positive qui reflète le fait remarquable qu'il existe un facteur de forme pour les consonnes s, z et ʃ, ʒ (figure 4).

4. Influence du contexte vocalique. L'influence régressive de la voyelle sur la consonne :

---

*. Nous entendons par là une classification à 100% sur les valeurs relevées.

- pour ʃ, ʒ : ne permet de séparer en moyenne (et encore à $\pm \sigma$) que les ʃ , ʒ suivis de voyelle labiale des ʃ , ʒ suivis de voyelle non labiale pour B mais pas pour A ni F_1 (figures 5 et 6); la protrusion ne permettant de distinguer qu'entre y, ø et i mais pas e (figure 6).

- pour s, z : l'assimilation vocalique ne se connaît pas davantage sur l'écartement A et la protrusion F_1, mais là aussi seulement pour l'aperture B (figures 5 et 6).

Nous retrouverons plus loin cette importance de B pour les voyelles elles-mêmes.

Labialité vocalique

La figure 7 présente les tracés moyens pour [i] et [y] suivant les contextes consonantiques.

1. Discrimination sur l'aperture (i/e, y/ø). Aucun paramètre ne permet d'obtenir une décision sur l'aperture entre voyelles très fermées et mi-fermées, et ceci même pas avec B (figure 8).

2. Discrimination sur la labialité (i, e / y, ø). Contrairement à ce qui se passe pour les consonnes, il est possible de faire une partition à 100% sur les valeurs observées pour A et B (figure 8) :

- A, correspondant à la distinction classique *écarté/non écarté*. Comme F_1 (soit *protrus/non protrus*) ne permet pas cette distinction, il s'avère que la terminologie non moins classique *arrondi/non arrondi* devrait être comprise simplement comme l'inverse d'*écarté/non écarté*.

- B, les labiales étant plus fermées aux lèvres que les non labiales. Comme,nous venons de le voir, l'aperture aux lèvres n'est pas corrélée avec l'aperture intrabuccale (hauteur de la langue et/ou abaissement du maxillaire) - ceci bien entendu dans les limites de la classe des voyelles fermées /i, e, y, ø/ - on est en droit de penser que l'aperture aux lèvres doit pouvoir fonctionner, dans un système de production ou de lecture labiale, comme un indice du trait de hauteur vocalique qui, passé un certain seuil n'est plus porteur que de l'information du trait de labialité.

En outre, le paramètre A, permet (ce que B ne permet pas) d'opérer une différenciation sub-phonémique entre [i], [i̯] et [y] (si/ʃi / sy = ʃy). On est donc amené à considérer, qu'au niveau phonétique, le trait de labialité peut être discrétisé , en ce qui concerne son indice A, comme un trait scalaire à 3 valeurs (0, 1, 2). Les données nous permettent aussi de le présenter sous sa forme polaire (-1, 0, +1) : la voyelle [y] sera phonologiquement *arrondie* (au sens restreint à l'inverse d'*écartée*) et [i], [i̯] phonologiquement *non arrondies*, ce dernier terme recouvrant à la fois les réalisations *écartées* (-1) et *neutres* (0). A est ainsi le corrélat physiologique le plus commode à traiter puisque B, bien que permettant une partition reflétant directement la distinction phonologique, n'est utilisable de façon bi-univoque pour le trait de labialité qu'entre certains seuils d'aperture intrabuccale.

3. Relations entre les paramètres. On ne trouve pour les voyelles une corrélation significative (au seuil 0,01) que pour les paramètres avec lesquels celle-ci était physiologiquement prévisible soit F_1, F_2 et B, D (figure 9). Ainsi on peut avancer qu'une des différences essentielles qui séparent les labialités vocalique et consonantique est l'absence d'un facteur de forme pour les

voyelles. Il resterait bien entendu à montrer par une étude EMG que, dès le niveau périphérique, la mimique labiale consonantique est commandée de manière physiologiquement différente pour les consonnes et les voyelles.

4. Influence des consonnes sur les voyelles. L'assimilation progressive se manifeste sur A et ceci uniquement pour les voyelles non intrinsèquement labiales (figure 8).

Les règles d'assimilation consonne-voyelle pour le français seraient donc schématiquement les suivantes, pour les paramètres les plus significatifs A et B :

- l'assimilation est progressive pour A, la consonne assimilant la voyelle. Mais cette assimilation ne joue que si la voyelle est [- *lab*] . Dans ce cas l'écartement vocalique augmente si la consonne est [- *lab*] et au contraire diminue si celle-ci est [+ *lab*] .

- l'assimilation est régressive pour B, la voyelle assimilant la consonne. L'aperture de celle-ci, quelle que soit la valeur du trait consonantique, augmente si la voyelle est [- *lab*], elle diminue si la voyelle est [+ *lab*].

CONCLUSION

En résumé, les caractéristiques les plus importantes de la labialité de mode consonantique et vocalique pour un locuteur francophone nous semblent être les suivantes :

- Pour les consonnes comme pour les voyelles, l'écartement A et l'aperture B sont les paramètres les plus significatifs, bien plus que la protrusion F et cela contre toute attente.

- Pour différencier les consonnes, quel que soit le contexte, deux paramètres sont nécessaires. A et B étant en outre significativement corrélés - et ceci seulement pour les consonnes - la présence d'un facteur de forme différencie la labialité consonantique de la labialité vocalique.

- Pour les voyelles, par contre, un seul paramètre suffit :
  . B reflète directement au niveau phonétique la distinction phonologique binaire + *lab* / - *lab*. Mais, ne l'oublions pas, l'aperture labiale étant corrélée à l'aperture intrabuccale, cette distinction risque fort de n'être utilisable que pour des degrés vocaliques voisins (i, e / y, ø).
  . A permet en plus pour des voyelles [- *lab*] de différencier, selon le contexte consonantique, [i̯, e̯] de [i, e]. C'est un corrélat non binaire à trois valeurs : *arrondi* [y, ø], *neutre* [i̯, e̯], *écarté* [i, e] dont les deux dernières réalisent la valeur phonologique négative du trait *arrondi*.

- Les règles de coarticulation consonne - voyelle nous donnent une assimilation progressive pour A, régressive pour B. La première a un domaine d'application restreint aux voyelles [- *lab*], la seconde est généralisée.

Il nous reste pour les mesures statiques à étendre ces conclusions aux autres locuteurs en y intégrant les mesures de l'aire C.

Les étapes suivantes consisteront à :

- interpréter ces résultats au niveau acoustique;
- passer aux mesures dynamiques;
- réaliser une expérimentation EMG complémentaire pour accéder au niveau des commandes.

REMERCIEMENTS

Nous tenons à remercier tout spécialement l'équipe cinématographique du S.I.R.P. (relations publiques des PTT) qui nous a permis, par sa compétence et aussi par sa gentillesse, de réaliser les films qui ont servi de base à nos études.

REFERENCES

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

FROMKIN, V., 1964, Lips positions in American English vowels; -Language & Speech 7, pp. 215-225.

LEANDERSON, R., 1972, On the functional organization of the facial muscles in speech; - Doct. Dissertation. Dept. of O.R.L. and Clinical Neurophysiology, Karolinska Sjukhuset - Stockholm.

LINDBLOM, B., 1965, Jaw-independence of labial parameters and a measure of labialization; -STL-QPSR 3, pp. 12-15.

LINDBLOM, B., 1965, Analysis of labial movement, STL-QPSR 2, pp. 20-25.

LINDBLOM, B., 1968, Studies of labial articulation; -Z. Phonetik Sprachwiss. Kommunikationforsch. 21, pp. 171-172.

LINDBLOM, B. & SORON, H., 1965, Analysis of labial movement; -J.A.S.A. 38, 935 (A).

LINDBLOM, B. & SUNDBERG, J., 1971, Acoustical consequences of lip, tongue, jaw and larynx movement; -J.A.S.A. 50, pp 1166-1179.

MAEDA, S.,1972, On the conversion of vocal tract X-ray data into formant frequencies; -Bell Laboratories - Memorandum for File MM72 - 1228 -6 .

ROBERTS, M., 1972, A cinephotographic study of coarticulation of lip rounding in English and French; -M. Sc. Dissertation University of British Columbia, Vancouver.

ROUSSELOT, P., 1923, Principes de phonétique expérimentale; -Tome I, H. Didier, Paris.

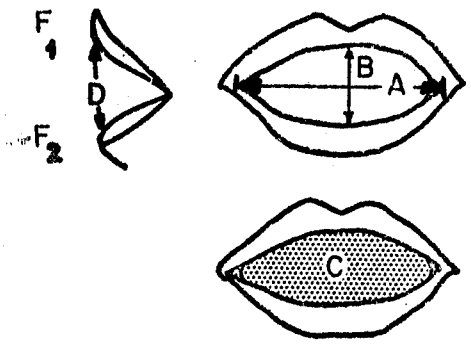


Figure 1 - Paramètres de la labialité; les protrusions F_1 et F_2 sont mesurées par rapport à la position préphonatoire.

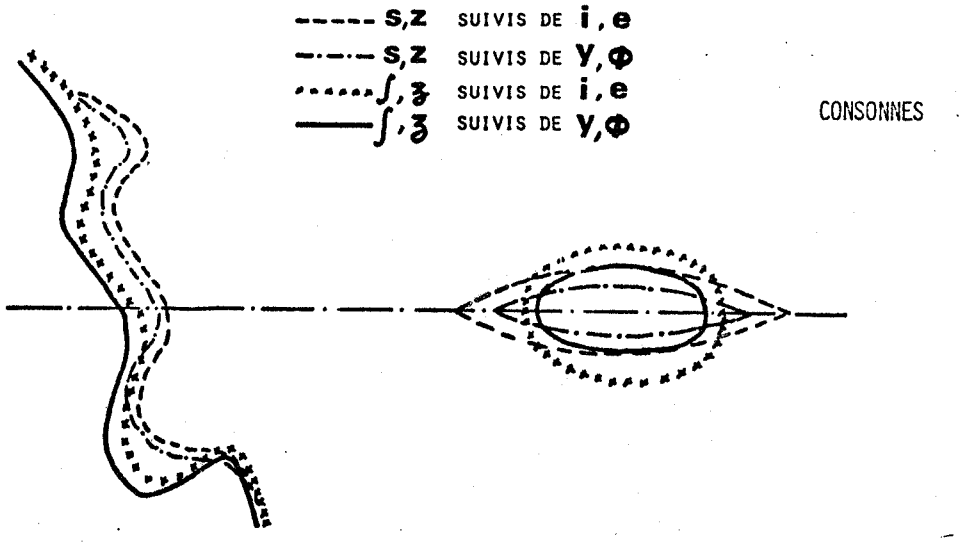


Figure 2 - Configurations labiales moyennes pour [s,z] [ʃ,ʒ] en fonction de l'entourage vocalique.

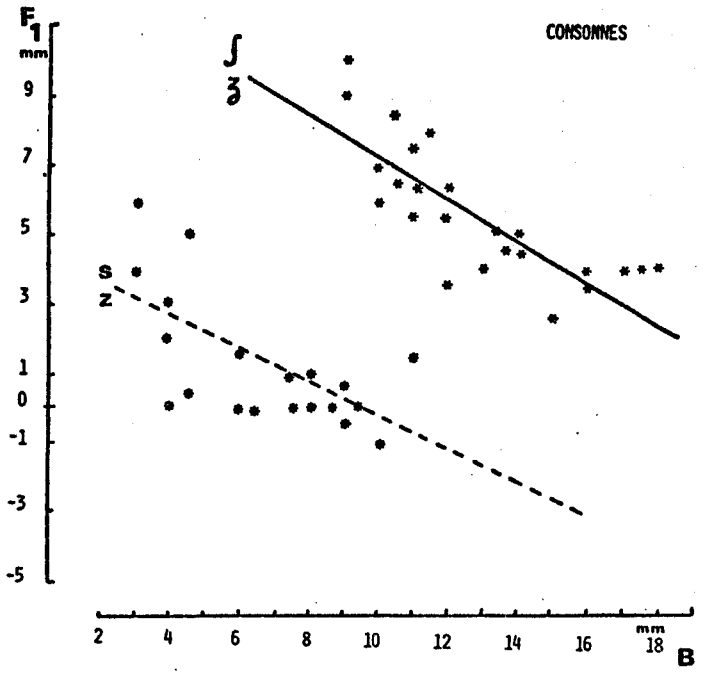


Figure 3 - Relations entre B et F_1, pour les consonnes, tous contextes vocaliques confondus.

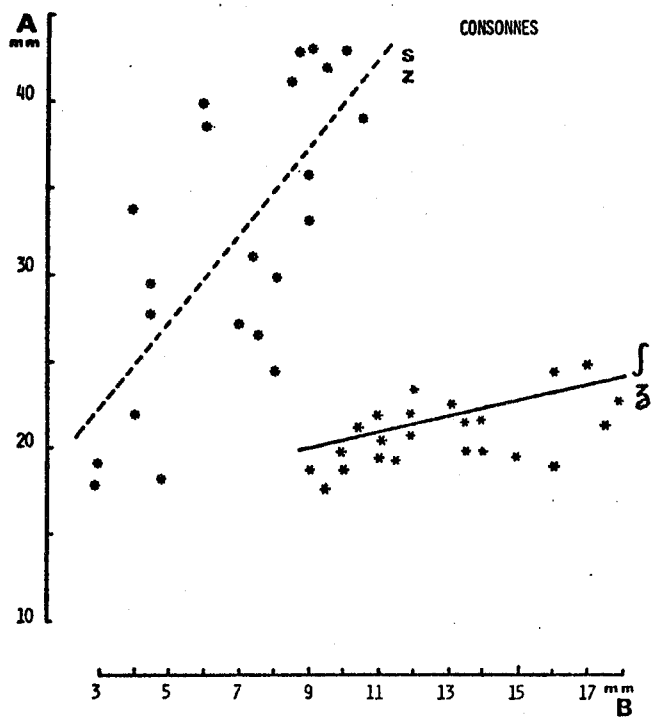


Figure 4 - Relations entre A et B, pour les consonnes, tous contextes vocaliques confondus.

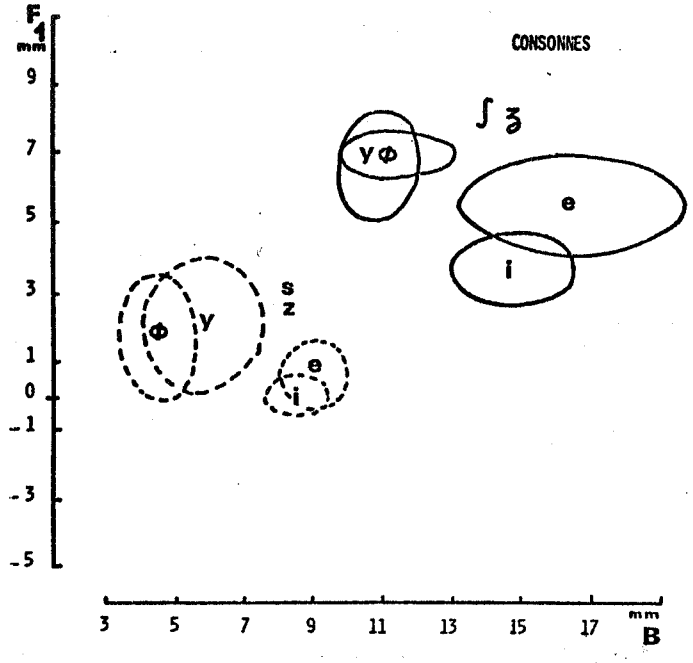


Figure 5 - Zones de dispersion F_1, B ($\bar{a} \pm \sigma$), pour les consonnes, selon les contextes vocaliques.

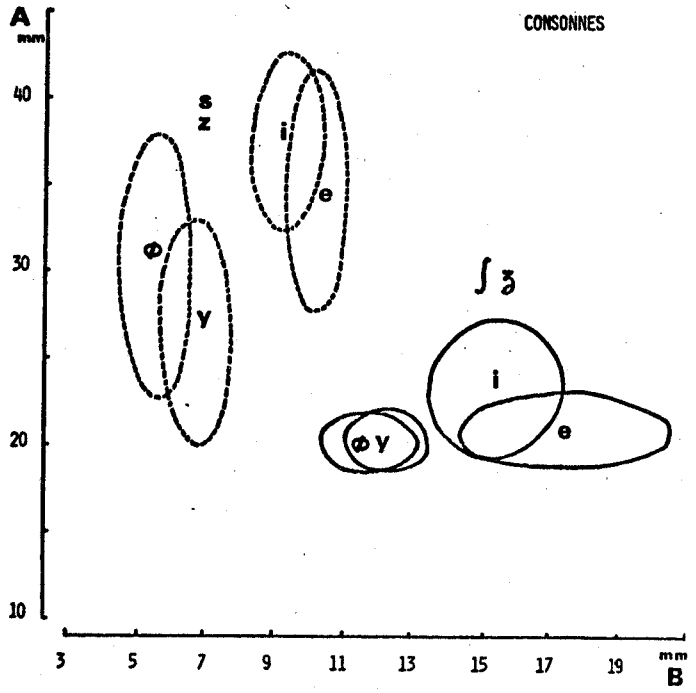


Figure 6 - Zones de dispersion A, B ($\tilde{a} \pm \sigma$) pour les consonnes selon les contextes vocaliques.

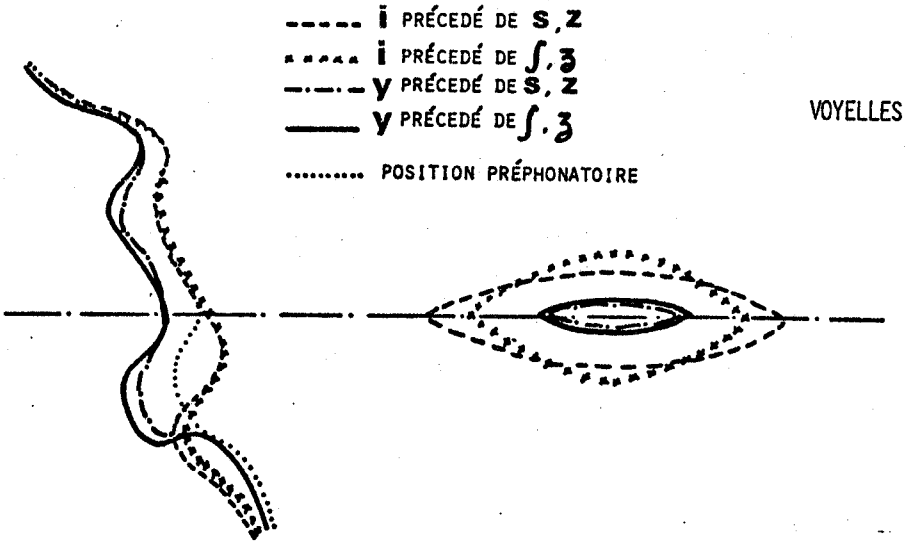


Figure 7 - Configurations labiales moyennes pour [i] et [y] selon les entourages consonantiques.

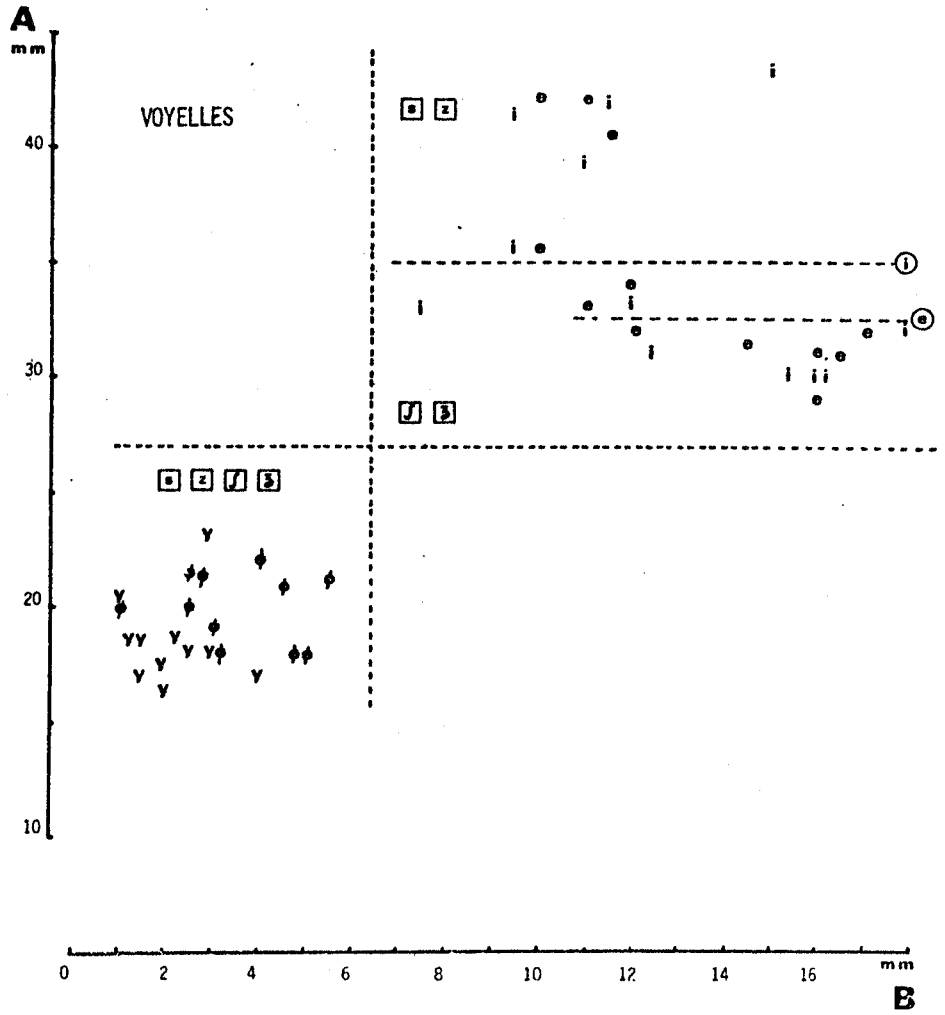


Figure 8 - Les classes de voyelles en A, B selon l'entourage consonantique - Sur A, la séparation en fonction du contexte consonantique est différente pour ⓘ et ⓔ .

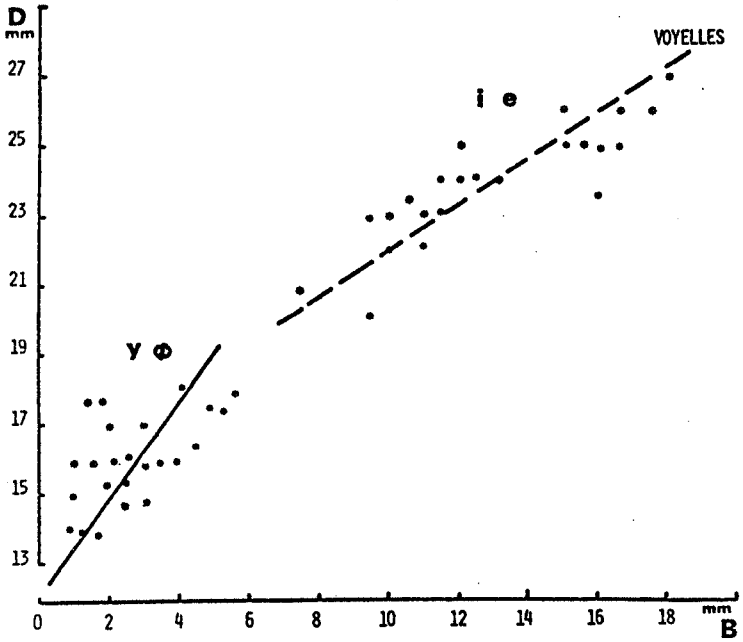


Figure 9 -

Relations entre D et B pour les deux classes de voyelles, tous contextes consonantiques confondus.

9èmes JOURNEES D'ETUDE SUR LA PAROLE

LANNION 31 mai - 2 juin 1978

UNE ANALYSE STATISTIQUE SUR LES POSITIONS DE LA LANGUE : ETUDE PRELIMINAIRE SUR LES VOYELLES FRANCAISES :

SHINJI MAEDA CNET - LANNION

.**RESUME**

Cet article traite d'une analyse en composantes principales du contour des coupes sagittales de la langue, pour 12 voyelles françaises, en contexte, prononcées par un locuteur.
Cette étude vise à apporter des éléments pour la construction d'un modèle articulatoire. 77 representations du conduit vocal, tirées d'un film radiocinématographique et publiées dans Brichler-Labaeye (1970) ont été retracées sur une table traçante interactive reliée à un ordinateur. L'état de chaque contour de la langue est mesuré à l'aide d'une coordonnée semi-polaire, (Voir Fig. 2) et est spécifié par 15 variables (Voir Fig. 3).

L'analyse a montré que les deux premières principales composantes rendent compte de 90 % de la variance des données originales. Si on utilise les trois premières composantes, le pourcentage s'élève à 98 %.

Le calcul de la racine de l'erreur quadratique moyenne (valeur efficace) dans l'estimation de chaque position de langue utilisant les composantes a été également effectué en fonction du nombre de ces composantes. La courbe de diminution du pourcentage d'erreurs (Voir Fig. 4) est différente d'une voyelle à l'autre.

On peut noter cependant que, lorsque les trois premières composantes sont utilisées, le taux d'erreurs descend en dessous de 3 %, et qu'il n'y a pas de différence significative de ce pourcentage entre les voyelles. En conséquence, on peut donc conclure que les trois composantes sont nécessaires et qu'elles suffisent peut-être pour décrire de façon adéquate les configurations de la langue pour les voyelles. L'article se termine par une discussion de l'interprétation à donner à ces composantes sur le plan articulatoire (voir Fig. 5).

A STATISTICAL ANALYSIS OF TONGUE SHAPES : A PRELIMINARY STUDY ON FRENCH VOWELS.

SHINJI MAEDA CNET - LANNION

SUMMARY

Principal component analysis of the midsagittal tongue outlines of 12 different French vowels uttered by a single speaker is described. This study aims to obtain some fundation for constructing an articulatory model.

Seventy seven x-ray tracings of the vocal tract outlines, which are published in Brichler-Labaeye (1970), were retraced on a computer interactive graphic table, and transfered into the memory (see Fig. 1). The state of each tongue outline was measured using a semipolar coordinate (see Fig 2) and was specified by 15 variables (see Fig. 3).

The analysis has shown that the first 2 principal components account for 90 % of the original data variance. If the first 3 components are used, the percentage increases to 98 %. The root mean square error in the estimation of each tongue outline was calulated also as a function of the number of the principal components. The manner of decrease in the error differs from vowel to vowel, when the number of the components is increased (see Fig. 4). The first 3 components are used, however, the errors become small, typically less than 3 %, and there is no significant difference in the amount of the error depending on the vowel identities. Thus it may be concluded that the first 3 components are necessary and perhaps sufficent to describe adequately the tongue outlines for the vowels.

Interpretation of the components is discussed in some extent. It appears that the first component represents a forward-up or backward-down tongue movement (see Fig. 5 (a), while the second component exhibits a tongue deformation as if its boby is pulled in backwar-up direction (see Fig. 5(b).). The third component seems to account a change in the size of the tongue body in the midsagittal section (see Fig. 5(c).).

INTRODUCTION

On peut décrire la parole sur le plan articulatoire ou sur le plan acoustique. Cependant les phénomènes de coarticulation par exemple, peuvent être décrits de façon plus efficace sur le plan articulatoire, ce qui peut justifier l'emploi d'un modèle articulatoire pour la synthèse de la parole. Notre analyse statistique sur les positions de la langue a pour but d'obtenir des données pour l'élaboration d'un tel modèle.

Un certain nombre d'auteurs ont récemment décrit une analyse statistique à plusieurs variables des états du conduit vocal. Une analyse des facteurs a été entreprise pour les voyelles anglo-américaines par HARSHMAN, R. (1970) et par HARSHMAN, R. & al. (1974 et 1976) ; OHTA, K. & al. (1974 et 1976) et KIRITANI, S. & al. (1976 et 1977) ont essayé de faire une analyse en composantes principales pour le japonais. Dans notre étude préliminaire, nous avons utilisé également une analyse en composantes principales ; cette méthode est simple et elle permet de trouver une solution élégante à la détermination du nombre minimal de paramètres nécessaires pour représenter les états avec une certaine précision connue pour les données utilisées.

Dans cet article, nous allons décrire l'analyse en composantes principales de la représentation des coupes sagittales de la langue pour 12 voyelles françaises, prononcées par un seul locuteur.

MATERIAU EXPERIMENTAL

Pour l'analyse nous avons utilisé les données publiées par BRICHLER - LABAEYE, C. (1970). Ces données consistent en des croquis de la coupe sagittale du conduit vocal à partir de films radio-cinématographiques enregistrés à une vitesse de 50 images/seconde. Les tracés ne représentent pas la région laryngale. Ils sont cependant suffisants pour étudier les positions de la langue, qui est l'organe articulatoire le plus important.

Nous avons choisi les données correspondant à 12 voyelles différentes. Le nombre de tracés pour une voyelle varie entre 3 et 11, et il dépend de sa durée. On peut noter que ces voyelles apparaissent dans des contextes phonétiques différents. Les 77 tracés ont été soumis à l'analyse en composantes principales.

MESURES SUR LES POSITIONS DE LA LANGUE

Ces mesures sont effectuées en deux étapes :

1. TRACE DES COUPES SAGITTALES DU CONDUIT VOCAL

Les contours tirés du film radio-cinématographique ont été retracés sur une tablette d'entrée graphique reliée à un ordinateur, afin de transférer les valeurs des coordonnées x-y des points décrivant ces contours dans une mémoire à disque de façon séquentielle. Pour le tracé, nous avons divisé le contour global du conduit vocal en 8 segments. Un exemple de tracé "segmenté" est illustré sur la Fig. 1, représenté sur la page suivante.

.../...

Les 8 caractères "p" indiquent le point terminal de chaque segment. L'adresse de la mémoire correspondant à chaque point terminal est également relevée. Le fait de retracer ces 8 segments dans un ordre donné permettra d'identifier chaque segment lors du traitement ultérieur, par exemple le segment correspondant au dos de la langue. De plus, cette méthode nous permet de mesurer les mouvements de points particuliers du conduit vocal image après image.

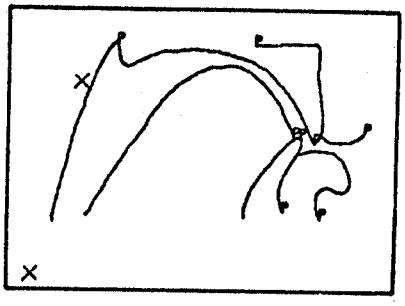


Fig. 1

Un exemple de tracé de la coupe sagittale de conduit vocal.

Dans la fig. 1, les deux signes "x" indiquent les points de référence choisis, qui ont été déterminés par rapport à la position du palais dur représenté dans les tracés originaux. Le calculateur effectue une projection du contour initial sur un système de coordonnées semi-polaires par rapport à ces deux points de référence (Voir Fig. 2).

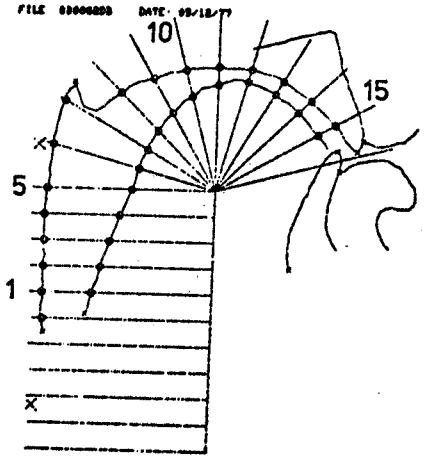


Fig. 2

La projection de ce tracé sur la coordonnée semi-polaire. Les cercles indiquent les intersections détectées. (Voir le texte).

Cette nouvelle projection des données originales fait appel à des opérations de rotation, translation et changement d'échelle. De cette façon, le traçé du contour sur la tablette d'entrée graphique peut se faire dans n'importe quelle direction, position et avec n'importe quelle échelle.

2. OBSERVATION DES POSITIONS DE LA LANGUE

Pour réaliser une analyse statistique, il est nécessaire de décrire les états à l'aide d'un nombre fixe de variables. Dans notre étude, les positions de la langue ont été mesurées à l'aide du système de coordonnées semi-polaires (Fig. 2). La grille de mesure est composée, dans la section linéaire, de lignes horizontales équidistantes de 1 cm environ et dans la section semi-circulaire, de rayons formant des angles de 15°.

Une procédure automatique basée sur un algorithme de recherche binaire détecte les points d'intersection entre la grille de mesure et le tracé du contour de la langue.

.../...

Sur la Fig. 2, les cercles représentent les intersections détectées(sur cette figure, les intersections de la paroi extérieure du conduit vocal sont également représentées). Les valeurs correspondantes des points d'intersection sont mesurées sur la grille de coordonnées semi-polaires.
Ces valeurs, qui caractérisent donc la forme de la langue, sont illustrées sur la Fig. 3.

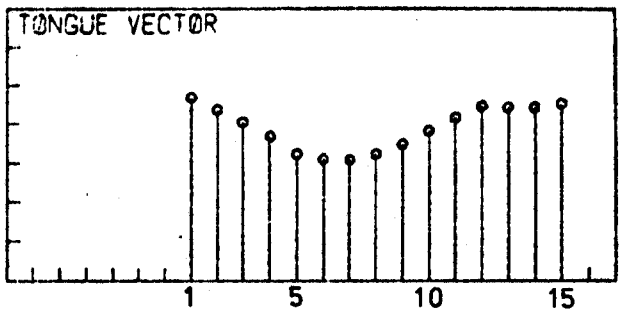


Fig. 3

Un vecteur composé de 15 variables qui décrivent le contour de la langue représentée sur la Fig. 2.

Nous avons indiqué par leur numéro les lignes de coordonnées (Fig. 2), et les variables correspondantes (Fig. 3).

On doit noter que le nombre de variables varie légèrement d'une image à l'autre, car la position de la base de la langue et de son extrémité varient pendant l'articulation. Afin d'obtenir un nombre fixe de variables pour les 77 tracés, nous avons simplement éliminé les variables absentes dans certaines projections. Après cette élimination, le nombre de variables s'est trouvé être de 15 pour nos données.

ANALYSE EN COMPOSANTES PRINCIPALES ET RESULTATS

Chaque tracé de la langue est décrit par 15 variables.
Ces variables sont :

x_i ou $i = 1,2,..p$ $(p = 15)$

qui forment un vecteur,

$X = (x_1\ x_2 \ldots\ldots x_p)^T$ .Dans la méthode d'analyse, la matrice de corrélation de X est estimée sur la base de 77 observations.

Les valeurs propres et les vecteurs propres de cette matrice sont calculés numériquement par la méthode de Jacobi (GREENSTADT, J., 1965). Les valeurs propres calculées, et les vecteurs propres qui leur correspondent sont alors ordonnés par valeurs décroissantes.

En construisant une matrice, Q, de telle façon que la K- ième colonne corresponde au K-ième vecteur propre, on peut obtenir une transformation linéaire du vecteur observé, comme suit :

$$Y = Q^T (X - \bar{X}), \quad (1)$$

où $\bar{X}$ représente le vecteur moyen de X.

Les éléments du vecteur transformé Y, c'est-à-dire,

y_k, où $k = 1,2,\ldots\ldots p$, sont appelés les composantes principales. Comme la matrice Q est orthogonale, la transformation inverse est donnée par :

$$X = QY + X. \quad (2)$$

.../...

La seconde équation est considérée comme une estimation par régression linéaire du contour de la langue, où les variables sont les composantes principales et les coefficients sont les éléments des vecteurs propres ordonnés.

Le fait remarquable des composantes principales est le suivant : alors que la série complète des p composantes peut représenter exactement le vecteur original X, (c'est-à-dire la forme exacte de la langue), il est possible de ne garder pour une approximation donnée que les n premières composantes ($n < p$), en sachant que ces n facteurs expliquent plus la variance de X que n'importe quelle autre série de n facteurs orthogonaux.

Notre analyse montre que les 2 ou 3 premières composantes suffisent à rendre compte d'une grande partie de la variance dans les données concernant la position de la langue. Le taux de distribution cumulée indique que les deux premières composantes rendent compte à elles seules de 90 % de la variance des données. Quand les trois premières composantes sont retenues, ce taux s'élève à 98 %.

Le taux de distribution cumulé peut être considéré comme une mesure approximative de la validité d'une estimation utilisant un nombre réduit de composantes principales. Ce taux n'indique pas, en conséquence, la précision avec laquelle est estimée chaque position individuelle de langue. A cette fin, nous avons calculé la racine de l'erreur quadratique moyenne faite dans l'estimation de chacun des contours observés. L'estimation de la i-ième variable avec les n premières composantes,

$$\hat{x}_i(n) = \sum_{k=1}^{n} v_{ik}\, y_k + \bar{x}_i \,, \qquad (3)$$

où v_{ik} = le i-ième élément du k-ième vecteur propre,

y_k = la k-ième composante principale,

$\bar{x}_i$ = la valeur moyenne de la i-ième variable.

On calcule alors la racine de l'erreur quadratique moyenne sous la forme d'un pourcentage, e (n) :

$$e(n) = 100 \times \sqrt{\sum_{i=1}^{p} \left(x_i - \hat{x}_i(n)\right)^2 \Big/ \sum_{i=1}^{p} x_i^2} \qquad (4)$$

Sur la Fig. 4, (sur la page suivante) les valeurs de e(n) sont indiquées en fonction de n, pour 9 différentes voyelles (non-labialisée). Chaque échantillon est sélectionné dans la partie médiane des voyelles. Le taux de diminution de l'erreur e(n) quand n augmente varie d'une voyelle à l'autre. Il apparaît cependant, que pour les 9 voyelles cette erreur descend en dessous de 3 % quand les trois premières composantes principales sont prises en compte pour l'estimation et qu'il n'y a pas de différence significative entre les taux d'erreur pour les différentes voyelles. Cela signifie que les contours de la langue pour les 9 voyelles peuvent être décrits avec une précision raisonnable par les trois premières composantes principales. Ce résultat est en accord avec celui qui a été obtenu pour l'analyse des voyelles anglo-américaines, où l'on montre que les trois premières composantes principales sont nécessaires pour rendre compte de plus de 95 % de la variance dans les données concernant la position de la langue (SHIRAI, K. & al ; 1976).

.../...

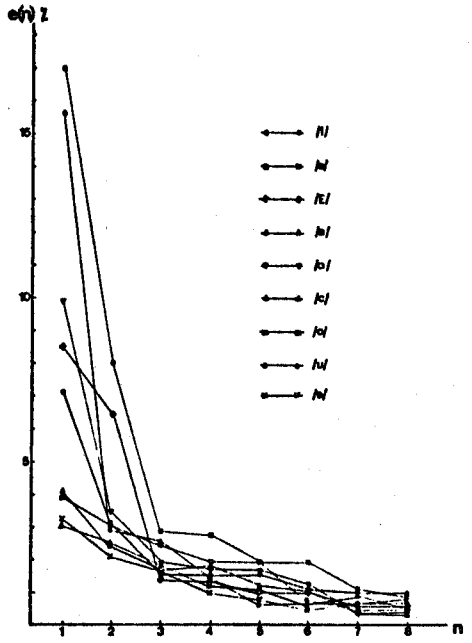


Fig. 4

Racine de l'erreur quadratique moyenne normalisée dans l'estimation d'un contour de la langue, à l'aide des n premières composantes principales, (e(n) est définie par l'équation (4)).

DISCUSSION ET CONCLUSION

De l'analyse décrite ci-dessus, on pourrait conclure que les 3 premières composantes principales sont nécessaires et probablement suffisantes pour décrire les contours de la langue pour les voyelles. La régression linéaire représentée par l'équation (3) peut donc être appliquée à la description d'un modèle articulatoire.

Pour une telle application, il peut être intéressant de rechercher la contribution de chaque composante principale à la position de la langue. L'influence de chacune d'entre elle peut être évaluée en terme d'écart par rapport au contour moyen de la langue. La méthode est la suivante :
On calcule $\hat{x}_i$ pour i = 1, 2,....p, en utilisant uniquement une composante principale, et on trace les valeurs trouvées sur le sytème de coordonnées semi-polaires.

Les résultats d'un tel calcul sont illustrés sur les figures 5a à 5c, pour chacune des tois premières composantes. Les valeurs des composantes indiquées à la fin de chaque contour, sont approximativement les valeurs maximales et minimales calculées lors de l'analyse. Le contour associé avec "o" dans chaque figure représente le contour moyen de la langue.

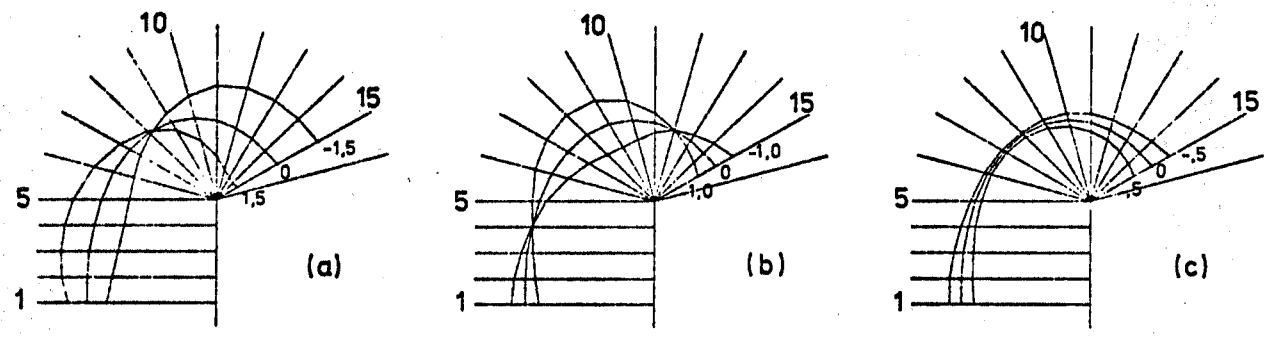


Fig. 5 Déplacement du contour de la langue à partir de sa position moyenne, qui est marquée par "0", dû à chacune des 3 premières composantes. Le mouvement dû à la 1-ière composante est représenté dans (a), à la 2-ième dans (b), et à la 3-ième dans (c).

.../...

La 1ère composante (dans la Fig. 5.a) montre un mouvement avant - montant, ou arrière - descendant. De tels mouvements font probablement intervenir deux facteurs : la position de la mâchoire et la déformation de la langue elle-même.

Les contours résultant de la 2-ième composante (Fig. 5b) montrent une contraction de la masse de la langue dans une direction arrière - montante. Les muscles extrinsèques de la langue, (probablement le styloglossus) participent, peut-être, en partie, à un tel mouvement.

La 3ième composante (Fig. 5c) montre une extension ou une compression de la coupe sagittale de la masse de la langue. Aucun des articles donnés en référence ne mentionne un tel mouvement. Nous ne somme pas en mesure, dès maintenant, de décider si une telle composante existe réellement dans l'articulation des voyelles françaises, ou si ce fait est simplement dû à la méthode utilisée pour les mesures.

Nous venons donc de donner une certaine interprétation des composantes principales. Nous admettons cependant qu'il est nécessaire de poursuivre des recherches sur leur nature d'un point de vue dynamique pour la construction d'un modèle articulatoire complet.

Par exemple, la position de la mâchoire a une influence certaine sur la forme de la langue. On peut en conséquence envisager une analyse dans laquelle l'influence due à la position de la mâchoire est déterminée. L'analyse en composantes principales est alors appliquée à la forme de la langue qui résulte après la suppression de l'influence de la position de la mâchoire. Kiritani (1977) a fait une estimation de la position de la mâchoire par une régression linéaire lors d'une telle analyse.
Nous avons essayé cette méthode pour nos données. Le degré d'ouverture de la mâchoire est calculé sur la base de la distance entre les incisives supérieures et inférieures. Les résultats indiquent que, pour obtenir une précision satisfaisante dans la reconstruction des contours de langue, il était nécessaire d'utiliser l'influence de l'ouverture de la mâchoire et les trois principales composantes. Cela signifie que 4 paramètres sont nécessaires pour une description adéquate des formes de la langue, au lieu des 3 utilisées dans l'analyse directe faite précédemment.

Cette augmentation du nombre de paramètres est, au moins en partie, due au fait que l'influence de la position de la mâchoire et les principales composantes de la forme de la langue résiduelle sont corrélées. En d'autres termes, la description par 4 paramètres contient une certaine redondance au sens statistique.

Afin de régler le problème de la redondance, nous avons commencé une analyse de facteurs dans laquelle le facteur dû à la position de la mâchoire n'est corrélé avec aucune des composantes principales. Nous espérons qu'une telle analyse conduira à des descriptions meilleures des contours de langue.

BIBLIOGRAPHIE

BRICHLER-LABAEYE, C. 1970, les voyelles françaises, mouvements et positions articulatoires à la lumière de la radiocinématographie, Klincksieck, Paris.

GREENSTADT, J., 1965, "Détermination des valeurs propres d'une matrice par la méthode de Jacobi", dans Méthodes Mathématiques pour Calculateurs Arithmétiques, A. Ralston et H. S. Wilf (eds) Ch. 7, pp 135 - 151 (traduit par P. Bévard), Dunod, Paris.

HARSHMAN, R., 1970, "PARFAC", Working Papers in Phonetics, University of California, Los Angeles, 16.

HARSHMAN, R. LADEFOGED, P., and GOLDSTEIN, L. 1977, "Factor analysis of tongue shapes", Journ. Acoust. Soc. Am., Vol. 62, N°3, pp. 693-707.

KIRITANI, S., and IMAGAWA, H., 1976, "Principal Component Analysis of Tongue Pellet Movement, "Annual Bulletin, Research Institute of Logopedics and Phoniatrics, University of Tokyo, 10, 15-17.

KIRITANI, S., SEKIMOTO, S., IMAGARWA, H., and FUJISAKI, H., 1977, "Parameter Description of the Tongue Movements for Vowels", Contribution papers of 9th ICA, Madrid, Vol 1,I13, 419.

OHTA, K., and Fuchi, 1974, "Multivariate Statistical Analysis of Estimated Vocal Tract Cross-sectional Area Function (1)", Progress on Speech Research '73, "Electrotechnical Laboratory, Tokyo 23-26.

OHTA, K., and IWAMATSU, A., 1976, "A Linear Articulatory Model and an Observation of Articulatory Movement of Semi-vowels based on the Estimated Vocal Tract Cross-sectional Area Function", Report at Group of Speech, Acoust. Soc. Japan, S76-05, (In Japanese).

SHIRAI, K., and HONDA, M., 1974, "An Articulatory Model and the Estimation of Articulatory Parameters", Report at Group of Speech, Acoust. Soc. Japan, S74-19 (in Japanese).

SHIRAI, K., and HONDA, M., 1976. "Estimations of Articulatory Motion", Seminar Records of U.S - Japan Joint Seminar on Dynamic Aspect of Speech Production, pp 83-96. (for information, contact with Prof. M. SAWASHIMA, Research Institute of Logopedics Phoniatrics, Univ. of Tokyo, JAPAN).

9èmes JOURNEES D'ETUDE SUR LA PAROLE

LANNION 31 mai - 2 juin 1978

UNE EXPLICATION SIMPLIFIEE, EN TERMES PHYSIQUES, DES CONSEQUENCES ACOUSTIQUES DES MOUVEMENTS DE LA LANGUE ET DES LEVRES DANS LA PRODUCTION DES VOYELLES

Nina Thorsen Institut de Phonétique, Université de Copenhague

.**RESUME**

Cet article est un extrait d'un exposé plus exhaustif qui essaie de fournir aux "non-ingenieurs" des notions fondamentales sur les relations entre articulation et acoustique des voyelles et des consonnes.

On peut expliquer, à l'aide de la seconde loi de NEWTON (force égale le produit de masse par accélération) et la loi de BOYLE-MARIOTTE (à température constante, le produit de pression par volume pour une certaine quantité d'air est constant) que "When a part of a pipe is constricted its resonance frequency becomes low or high according as the constricted part is near the maximum point of the volume current ... or of the excess pressure ..." [Quand une partie d'un tuyau est rétrécie la résonance est basse ou haute selon que le rétrécissement se trouve près du point de vélocité maximum ou près du point de pression maximum.] (CHIBA, T. & KAJIYAMA, M., 1958:151).

Ceci se démontre en prenant en considération les forces relatives qui opèrent sur une mince couche d'air qui oscille à l'ouverture d'un résonateur de type 'fermé-ouvert' ("quarter-wavelength" ["quart d'onde"]) à sa première résonance. Une diminution du volume du tuyau près de l'extrémité fermée entraîne un accroissement des forces qui entretiennent l'oscillation de l'air et, par conséquent, augmente sa fréquence, et vice versa. Au contraire, en diminuant l'ouverture du résonateur on fait décroître les forces qui entretiennent les oscillations et, par conséquent, la fréquence diminue, et vice versa.

A SIMPLIFIED EXPLANATION, IN PHYSICAL TERMS, OF THE ACOUSTICAL CONSEQUENCES OF TONGUE AND LIP MOVEMENT IN VOWEL PRODUCTION

Nina Thorsen Institute of Phonetics, University of Copenhagen
96, Njalsgade, DK 2300 Copenhagen, Denmark

SUMMARY [A full English translation can be obtained from the author.]

This paper attempts to explain, in a simple fashion, the often-cited fact that "When a part of a pipe is constricted its resonance frequency becomes low or high according as the constricted part is near the maximum point of the volume current ... or of the excess pressure ..." (CHIBA, T., & KAJIYAMA, M., 1958:151).

A number of simplifications are introduced: the vocal tract is regarded as a perfect quarter-wavelength resonator (fig.1); there is no loss of acoustic energy and no radiation from the lips. At the first resonance the oscillations of the column of air in the pipe can be compared with those of a mass and spring (fig.3). In the case of the vibrating air column, the equivalent of the mass is a thin slice of air, S, at the open end of the pipe, and the equivalent of the spring is the volume of air behind the opening (fig.4 - the vibratory patterns depicted in the figures are grossly simplified and exaggerated).

There is a close tie between the forces that act on S, S's mass and S's motion, which is given by NEWTON's second law [1.1]. In the case of the uniform pipe, this law can be paraphrased to say that S's acceleration, which is indicative of S's mean velocity, is proportional to the size of the pressure variation in the pipe, which arises from S's motion [1.2-5]. Assuming that the changes in volume, due to S's motion, are small compared to the total volume, BOYLE-MARIOTTE's law [2.1] tells us then that the pressure variations in the pipe are proportional to the volume changes [2.2-7], i.e. to the elongation of S (fig.5). From this follows that for a given volume of air, S's elongation is proportional to the pressure changes it induces on the volume, i.e. to the forces that keep it in motion, i.e. to its mean velocity [3.1-3], and thus its frequency is unaffected by changes in its elongation. Diminishing the volume behind S (fig.6) will increase the magnitude of the pressure variations induced in the pipe for any given elongation of S and thus increase the forces that keep S in motion, i.e. increase its mean velocity and thereby raise its frequency of oscillation.

The smaller the opening, the smaller the mass of the vibrating slice of air (fig.9; [4.1-2]) and, for any given pressure increment, the smaller the forces that act on S, since force is the product of pressure and the area of the surface on which the pressure applies [5.1-2]. Combining this fact with NEWTON's second law we see that for a given change of pressure the acceleration (mean velocity) of S_1 and S_2 (fig.9) is the same [6.1-3]. But their elongations are not! In order to induce in R_2 the same pressure increment as in R_1, S_2 must have twice the elongation of S_1, and thus its frequency is half that of S_1.

By considering the pipe as composed of three, five etc. shorter quarter-wavelength pipes, put together front-to front and back-to-back (figs.8, 11), one can reason about the second, third, etc., resonances in a similar way.

However, pressure and velocity variations are not concentrated at the closed and open ends of the pipe, but are distributed sinusoidally along the length of the pipe (figs.2, 8, 11), and a change in volume or an occlusion will therefore affect a given resonance more effectively, the closer it is to a pressure or a velocity maximum for that resonance. Further: due to the limitations imposed by the articulatory organs a constriction along the vocal tract will act as a volume change and an occlusion at the same time, which fact leads us to the general formulation, cited in the first paragraph.

UNE EXPLICATION SIMPLIFIEE, EN TERMES PHYSIQUES, DES CONSEQUENCES ACOUSTIQUES DES MOUVEMENTS DE LA LANGUE ET DES LEVRES DANS LA PRODUCTION DES VOYELLES

Nina Thorsen Institut de Phonétique, Université de Copenhague

INTRODUCTION

Cet exposé ne prétend pas être scientifique et original au sens habituel de ces mots. On essaye simplement d'expliquer aux phonéticiens sans formation spécifique en physique et en mathématique, plus simplement que ne le font la plupart des articles et des livres sur ce sujet, le fait, si souvent cité, que "When a part of a pipe is constricted its resonance frequency becomes low or high according as the constricted part is near the maximum point of the volume current ... or of the excess pressure ..." [Quand une partie d'un tuyau est rétrécie la résonance est basse ou haute selon que le rétrécissement se trouve près du point de vélocité maximum ou près du point de pression maximum.] (CHIBA, T. & KAJIYAMA, M., 1958:151).

SIMPLIFICATIONS PRELIMINAIRES

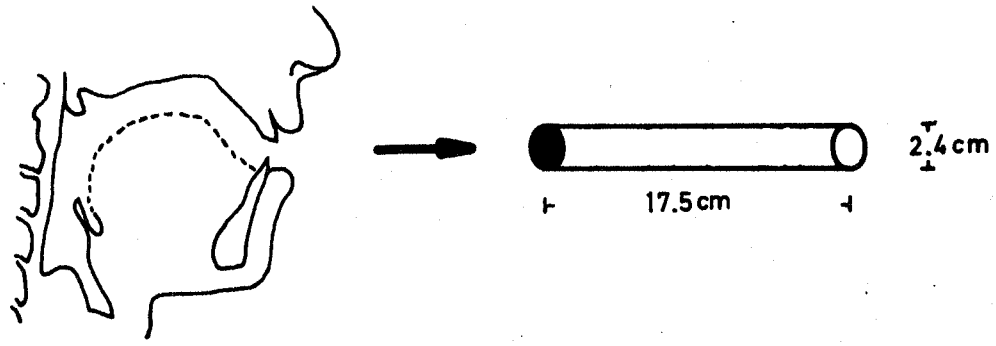


Figure 1

Le conduit vocal réel et un modèle simplifié du conduit vocal.

The vocal tract and a simplified model of the vocal tract.

Supposons que le conduit vocal soit un tuyau cylindrique, de longueur de 17,5cm et de diamètre de 2,4cm, fermé à une extrémité, ouvert à l'autre (fig.1). Admettons en outre que les parois du tuyau sont parfaitement rigides (c.-à-d. qu'elles n'absorbent pas d'énergie acoustique) et qu'il n'y a aucun rayonnement d'énergie de l'ouverture du tuyau vers l'extérieur (c.-à-d. qu'il n'y a pas de diffraction du son depuis les lèvres. C'est une absurdité monstrueuse, bien sûr, et en pratique cela voudrait dire que nous nous n'entendons pas parler, mais c'est une simplification opératoire et qui n'est pas, d'ailleurs, un obstacle sérieux pour les considérations qualitatives qui suivent). Ainsi, c'est un résonateur uniforme parfait, de type 'quarter-wavelength' ('quart d'onde'), les résonances duquel sont environ 500, 1500, 2500, .. Hz.

LE MODE D'OSCILLATION DANS LE TUYAU 'QUART D'ONDE' A SA PREMIERE RESONANCE

Comme point de départ prenons le tuyau uniforme (fig.1) (c.-à-d. la voyelle neutre). Observons la colonne d'air dans ce tuyau, mise en oscillation à sa première résonance, et supposons que cette oscillation continue sans décroissement d'amplitude tout le temps de l'observation. (En pratique c'est impossible sans apport constant d'énergie, ce qui n'a aucune importance pour cette "démonstration".)

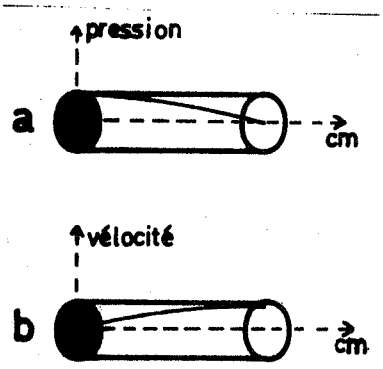


Figure 2

La distribution de pression (a) et de vélocité (b) dans un tuyau uniforme 'quart d'onde' à sa première résonance.

The distribution of pressure (a) and velocity (b) in a uniform quarter-wavelength pipe at its first resonance.

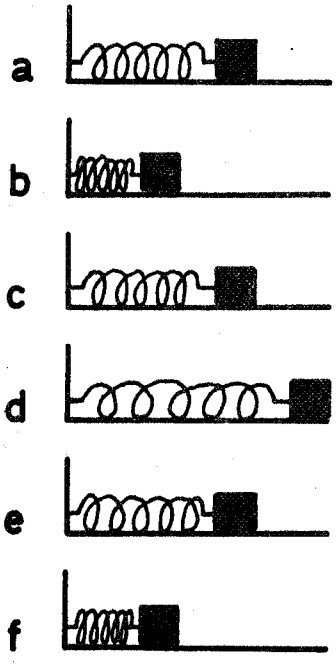


Figure 3

Le mode d'oscillation d'un poids, fixé à un mur rigide par un ressort, et qui glisse sur une surface lisse.

Vibratory pattern of a spring and mass, attached to a hard wall, sliding on a smooth surface.

Nous savons qu'à l'ouverture du tuyau (les lèvres) la variation de vélocité[1] est maximum, c.-à-d. que les particules d'air oscillent de part et d'autre de l'ouverture avec une élongation maximum. A l'extrémité fermée (la glotte) la variation de pression[2] est à son maximum. Nous savons de plus qu'il y a des variations de pression et de vélocité, respectivement, dans toute l'étendue du tuyau (fig.2) mais que la pression décroît de l'extrémité fermée à l'ouverture, où elle est nulle, et que la vélocité décroît de l'ouverture à l'extrémité fermée, où elle est nulle. (Ces faits aussi peuvent être expliqués, mais non pas sans dépasser les limites de cet article.)

Cependant, tant que l'on ne regarde que la première résonance, on peut admettre, pour la colonne d'air, que tout mouvement, c.-à-d. vélocité, se trouve concentré à l'ouverture et que toute variation de pression se trouve concentrée à l'extrémité fermée du tuyau. (Ainsi c'est un système à constantes concentrées, avec un degré de liberté, c.-à-d. qu'il ne peut osciller qu'à une seule fréquence.) Dans ce cas on peut comparer le système acoustique à un système méchanique, composé d'un poids, fixé à un mur rigide par un ressort, et qui glisse sur une surface parfaitement lisse, c.-à-d. qu'il n'y a pas de frottement entre le poids et la surface quand le poids oscille (fig.3).

Tout mouvement présuppose une force: si l'on déplace le poids à gauche (3b) la compression du ressort exerce une force à droite, et quand nous lâchons le poids cette force le met en mouvement à droite, vers la position d'équilibre. Le poids dépasse la position d'équilibre (3c), parce que tout corps qui a une masse a aussi de l'inertie, ce qui veut dire que le mouvement d'un corps continue après que la force qui l'a initié ait cessé d'opérer. Ainsi le ressort devient de plus en plus long et il exerce une force croissante à gauche. A un certain moment le mouvement à droite s'arrête (3d), et un mouvement à gauche commence, vers la position d'équilibre. A cause de son inertie, le poids va dépasser la position d'équilibre encore une fois (3e), le ressort est comprimé de nouveau, et il exerce une force croissante à droite, jusqu'au moment où le mouvement du poids est arrêté (3f) et un mouvement à droite commence, et ainsi de suite. S'il n'y a aucune perte d'énergie, le poids va osciller éternellement avec une élongation constante. Sa fréquence est déterminée par l'élasticité du ressort et la masse du poids. Plus l'élasticité du ressort est grande et plus la masse du

---

1) 'variation de vélocité' est parfois abrégé en 'vélocité' par la suite.
2) 'variation de pression' est parfois abrégé en 'pression' par la suite.

poids est petite, plus la fréquence sera élevée, et vice versa. L'élongation du poids ne dépend que du déplacement initiale.

De la même façon on peut observer le comportement d'une mince couche d'air, S, à l'extrémité ouverte du tuyau (fig.4) (les mouvements et l'épaisseur de ces couches d'air sont énormément exagérés dans les figures). Ce qui entretient les mouvements de cette couche d'air, c'est l'action conjugée (1) des variations de pression qui naissent dans le tuyau à cause des mouvements de S, et (2) de l'inertie de S. Si S commence son mouvement à droite (4a) c'est que la pression dans le tuyau est plus grande qu'à l'extérieur du tuyau (la pression atmosphérique), et si S dépasse la position d'équilibre (4b), c'est que S a une certaine masse (même si elle est très petite), et par conséquent de l'inertie. Ainsi la pression dans le tuyau décroît, c.-à-d. que la pression atmosphérique à l'extérieur du tuyau exerce une force croissante (par rapport à la pression dans le tuyau) à gauche, qui finalement arrête S et la met en mouvement à gauche (4c). Cette oscillation continue éternellement avec amplitude constante, s'il n'y a aucune perte d'énergie nulle part.

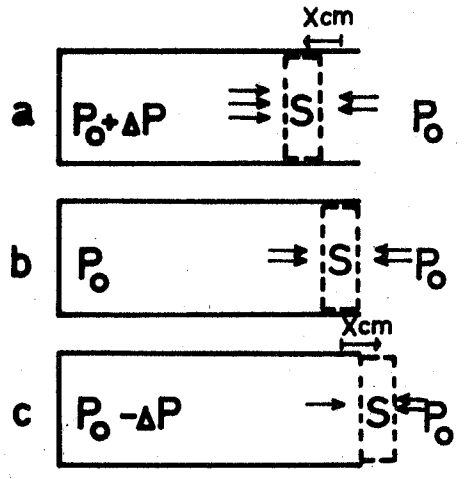


Figure 4

Le mode d'oscillation d'une mince couche d'air à l'extrémité ouverte du tuyau uniforme 'quart d'onde'.

Vibratory pattern of a thin slice of air at the open end of the uniform quarter-wavelength pipe.

Il y a une relation très étroite entre les forces qui opèrent sur S, la masse de S et le mouvement de S, qu'on peut exprimer dans une équation (la seconde loi de NEWTON):

[1.1] $F = m \cdot G$ (force égale le produit de masse par accélération)

a) la force, dans notre cas, est le produit de la pression par la surface sur laquelle opère la pression

[1.2] $F = P \cdot A$

Cette surface est constante (fig.4). Ce qui varie, c'est la pression dans le tuyau.

b) S est simultanément influencée par deux forces antagonistes, une qui est due à la pression dans le tuyau et une qui est due à la pression atmosphérique. La force résultante est due à la différence entre ces deux pressions.

c) la masse de S est constante.

Ainsi on peut paraphraser la seconde loi de NEWTON:

[1.3] $P \cdot A = m \cdot G$ d'où

[1.4] $((P_0+\Delta P) - P_0) \cdot A = m \cdot G$ donc

[1.5] $\Delta P = k \cdot G$

où P_0 est la pression atmosphérique, ΔP est l'accroissement (ou décroissement) de pression dans le tuyau, k est une constante qui est égale à la masse de S divisée par sa surface, et G est l'accélération de S, qu'on peut considerer comme une expression de la vitesse moyenne de S. C.-à-d. que la vitesse de S varie

en fonction de la différence entre la pression à l'intérieur du tuyau et à l'extérieur du tuyau. Cette différence est positive et negative, tour à tour, et ainsi S bouge d'un côté à l'autre à travers l'ouverture du tuyau.

TUYAUX NON-UNIFORMES A OUVERTURE CONSTANTE

On peut montrer que

a) l'élongation de S ne dépend que de la grandeur de la force initiale.

b) l'élongation et la fréquence de S sont indépendantes l'une de l'autre.

c) la fréquence de S dépend seulement de la grandeur relative des variations de pression qui naissent dans le tuyau à cause des mouvements de S, grandeur qui est déterminée par le volume total du tuyau uniforme.

Tout cela est une conséquence de la seconde loi de NEWTON et d'une autre loi qui dit que, à température constante, le produit de pression par volume pour une certaine quantité d'air est constant (la loi de BOYLE-MARIOTTE):

$$[2.1] \quad P \cdot V = k$$

Dans notre cas cela veut dire que si le volume de la colonne d'air est augmenté par le mouvement de S hors du tuyau, la pression dans le tuyau décroît, et vice versa. Le fait le plus important, dans ce cas, c'est que si les variations de volume sont petites par rapport au volume total, les variations de pression sont proportionelles aux variations de volume: Nous appellons P_0 et V_0 la pression et le volume, respectivement, en position d'équilibre, et obtenons ainsi:

$$[2.2] \quad P_0 \cdot V_0 = k$$

Nous diminuons le volume de ΔV et obtenons un accroissement de pression de ΔP_1:

$$[2.3] \quad (P_0+\Delta P_1)(V_0-\Delta V) = k, \quad \text{c.-à-d.}$$

$$[2.4] \quad \Delta P_1(V_0-\Delta V) = k - P_0(V_0-\Delta V) = P_0 \cdot V_0 - P_0(V_0-\Delta V) = P_0 \cdot \Delta V$$

Nous diminuons le volume de $2 \cdot \Delta V$ et obtenons ainsi un accroissement de pression de ΔP_2:

$$[2.5] \quad (P_0+\Delta P_2)(V_0-2\Delta V) = k, \quad \text{c.-à-d.}$$

$$[2.6] \quad \Delta P_2(V_0-2\Delta V) = k - P_0(V_0-2\Delta V) = P_0 \cdot V_0 - P_0(V_0-2\Delta V) = 2 \cdot P_0 \cdot \Delta V$$

donc $$[2.7] \quad \Delta P_2 = \frac{2 \cdot P_0 \cdot \Delta V}{V_0-2 \cdot \Delta V} = \frac{2 \cdot \Delta P_1(V_0-\Delta V)}{V_0-2 \cdot \Delta V} = 2 \cdot \Delta P_1 \quad \text{si} \quad \Delta V \ll V_0$$

C'est à dire que si une diminution de volume de $\Delta V cm^3$ cause un accroissement de pression de ΔPbaries, une diminution de $2 \cdot \Delta V cm^3$ donnera un accroissement de pression de $2 \cdot \Delta P$baries. (En pratique les variations de volume sont très petites, car l'élongation des particules d'air sont d'un ordre de grandeur de millionièmes d'un millimètre.)

Ad (a) et (b): Disons que la force initiale qui met S en mouvement est un déplacement de S de Xcm à gauche (fig.5a). Cela produit une pression de $(P_0+\Delta P)$ baries dans le tuyau. Quand nous lâchons S, elle se met en mouvement à droite, et on sait que l'accélération (la vitesse moyenne) de S est proportionelle à la différence entre la pression dans le tuyau et la pression atmosphérique:

$$[3.1] \quad (P_0+\Delta P) - P_0 = k \cdot G_1 \quad \text{d'ou} \quad G_1 = \frac{\Delta P}{k}$$

Si l'on commence par déplacer S de 2Xcm à gauche (5b), la pression sera

$(P_0+2\Delta P)$baries. L'accélération devient:

$$[3.2] \quad G_2 = \frac{(P_0+2\Delta P) - P_0}{k} = \frac{2\Delta P}{k}$$

donc $$[3.3] \quad G_2 = 2G_1$$

La vitesse moyenne de S dans le second cas est deux fois celle dans le premier cas, mais son élongation est, elle aussi, deux fois celle dans le premier cas, et ainsi la période et la fréquence sont égales dans les deux cas.

Ad (b) et (c): Si l'on déplace S_1 et S_2 de Xcm à gauche dans des tuyaux à volumes de Vcm^3 et $\frac{1}{2}Vcm^3$ (fig.6), le décroissement de volume relatif dans (6b) est deux fois celui dans (6a), et ensuite l'accroissement de pression dans (6b) est deux fois celui dans (6a). La force qui opère sur S_2 est deux fois celle qui opère sur S_1. La vitesse moyenne de S_2 est donc deux fois celle de S_1. Si leurs élongations sont égales, la fréquence de S_2 est deux fois celle de S_1. (Ceci est en accord avec les résultats qu'on obtient par les formules pour les résonances dans les tuyaux uniformes de type 'quart d'onde':

$$f_n = \frac{c}{4 \cdot L}(2n-1)$$

ou c est la célerité du son, L la longueur du tuyau et n le numéro de la résonance. Si c=35000cm/sec, L_1=17,5cm (a) et L_2=8,75cm (b) on obtient

$$(a): \quad f_1 = \frac{35000}{70} = 500\text{Hz}$$

$$(b): \quad f_1 = \frac{35000}{35} = 1000\text{Hz}\ .)$$

Le modèle d'oscillation décrit ci-dessus est extrêmement simplifié, car la vélocité et les variations de pression ne sont pas concentrées à l'ouverture et à l'extrémité fermée, respectivement, du tuyau, voir la fig. 2. En pratique, cela veut dire que les changements de volume produisent le plus grand effet sur la première résonance s'ils se trouvent près de l'extrémité fermée du tuyau, où les variations de pression sont à leur maximum.

On peut conclure que la première résonance du tuyau de la fig.7 en haut, qui est un modèle de la voyelle [ɑ], doit être plus haute que celle d'un tuyau uniforme ([ə]) et que, inversement, la première résonance du tuyau de la fig.7 en bas, qui est un modèle de la voyelle [i], est plus basse que celle de [ə], ce qui est confirmé par les faits empiriques. (Voir aussi le résumé.)

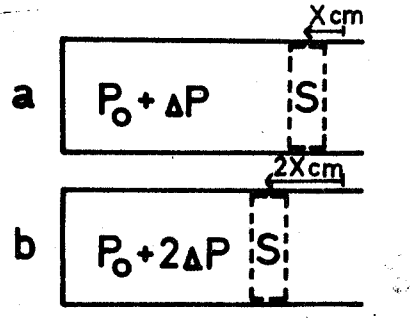


Figure 5

La relation entre l'élongation initiale de S et l'accroissement de pression dans le tuyau uniforme.

The relationship between initial elongation of S and pressure increment in the uniform pipe.

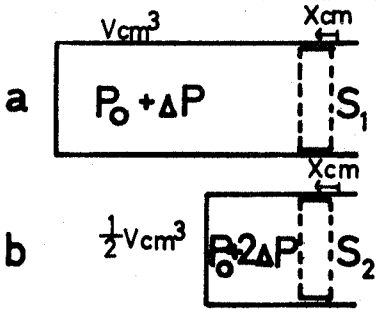


Figure 6

La relation entre l'accroissement de pression et l'élongation de S dans deux tuyaux uniformes à volumes Vcm^3 (a) et $\frac{1}{2}Vcm^3$ (b).

The relationship between pressure increment and elongation of S in two uniform pipes with volumes Vcm^3 (a) and $\frac{1}{2}Vcm^3$ (b).

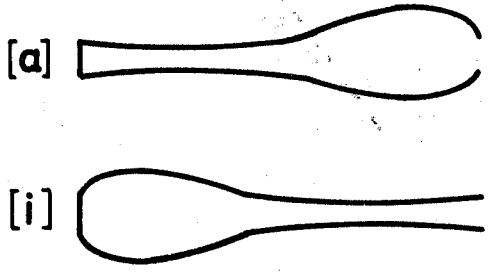


Figure 7

Modèles des deux voyelles [ɑ] et [i].

Models of the two vowels [ɑ] and [i].

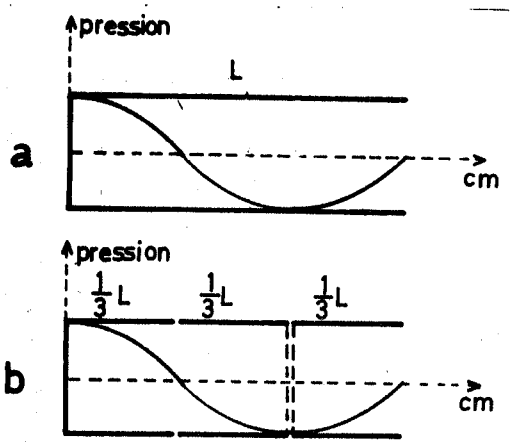


Figure 8

La distribution de pression dans un tuyau uniforme à sa seconde résonance.

The pressure distribution in a uniform pipe at its second resonance frequency.

Si l'on veut considérer l'effet des changements de volume sur les seconde, troisième etc. résonances, on ne peut plus comparer le système à une seule couche d'air (un poids) et un seul volume avec des variations de pression (un ressort). La distribution de pression le long du tuyau à sa seconde résonance est representée dans la fig.8. Si la colonne d'air n'oscille qu'à sa seconde résonance le systeme se comporte comme s'il était formé de trois tuyaux, chacun d'une longueur de 1/3 Lcm (fig. 8b). Les deux tuyaux imaginaires à gauche sont joints, ouverture contre ouverture, et le tuyau à droite est joint, bout fermé contre bout fermé, avec le tuyau au milieu. Maintenant on peut raisonner sur le système de la même façon que pour la première résonance, seulement il y a deux endroits où un changement de volume aura un effet appréciable sur la (seconde) résonance, à savoir à l'extrémité fermée et à une distance de 2/3 L de l'extrémité fermée. A la troisième résonance il y aura trois endroits, à la quatrième quatre endroits, etc., où un changement de volume aura un effet appréciable sur la fréquence. Chaque résonance a un maximum de pression à l'extrémité fermée du tuyau, et par conséquent chaque résonance monte ou baisse à la suite d'une constriction ou d'un élargissement près du bout fermé (mais pas toujours à un degré égal, voir le résumé).

TUYAUX UNIFORMES A OUVERTURE VARIABLE

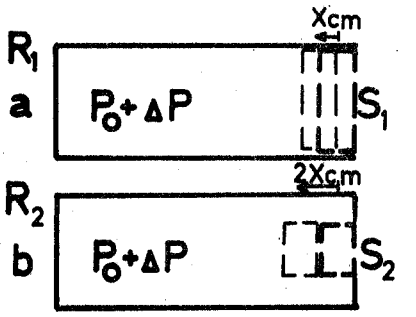


Figure 9

La relation entre l'élongation de deux couches d'air à surfaces Acm² (a) et ½Acm² (b) et l'accroissement de pression dans le tuyau uniforme.

The relationship between the elongation of two slices of air with areas Acm² (a) and ½Acm² (b) and pressure increment in the uniform pipe.

Voyons ce qui se passe si l'on diminue ou augmente le volume à l'extrémité ouverte du tuyau. Le changement de volume en soi n'a presque aucun effet puisque les variations de pression sont nulles à l'ouverture, mais l'occlusion verticale est essentielle, et c'est elle seule qui est representée dans la fig.9. On compare deux tuyaux uniformes de diamètres égaux et de longueurs égales. Le tuyau en haut, R_1, est tout ouvert, c.-à-d. que la surface de l'ouverture est Acm². Le tuyau en bas, R_2, a une ouverture circulaire de ½Acm². Les deux couches oscillantes, S_1 et S_2, sont de même épaisseur, Bcm. Utilisons les lois de NEWTON et de BOYLE-MARIOTTE: $F = m \cdot G$ et $P \cdot V = k$.

Les masses de S_1 et S_2 sont connues si l'on connaît le volume et la densité, ρ:

[4.1] $m_1 = A \cdot B \cdot \rho$ [4.2] $m_2 = \frac{1}{2}A \cdot B \cdot \rho$

Pour un certain changement de pression, ΔP, dans R_1 et R_2 on obtient des forces, F_1 et F_2, qui opèrent sur S_1 et S_2 de la manière suivante:

[5.1] $F_1 = \Delta P \cdot A$ [5.2] $F_2 = \Delta P \cdot \frac{1}{2}A$

Mais une force est aussi le produit d'une masse par une accélération, donc:

[6.1] $F_1 = \Delta P \cdot A = m_1 \cdot G_1$ d'où $G_1 = \frac{\Delta P \cdot A}{m_1} = \frac{\Delta P \cdot A}{A \cdot B \cdot \rho} = \frac{\Delta P}{B \cdot \rho}$

[6.2] $F_2 = \Delta P \cdot \frac{1}{2}A = m_2 \cdot G_2$ d'où $G_2 = \frac{\Delta P \cdot \frac{1}{2}A}{m_2} = \frac{\Delta P \cdot \frac{1}{2}A}{\frac{1}{2}A \cdot B \cdot \rho} = \frac{\Delta P}{B \cdot \rho}$ d'où

[6.3] $G_1 = G_2$

L'accélération (la vitesse moyenne) des deux couches sera donc la même. MAIS elles n'auront pas la même élongation: afin de donner à R_1 et R_2 la même diminution de volume, et donc le même accroissement de pression, on doit déplacer S_2 d'une distance qui est deux fois celle de S_1, parce que la surface (et le volume) de S_2 n'est que la moitié de celle de S_1. Si la vitesse moyenne des deux couches est égale et que l'une parcourt une distance qui est deux fois celle parcourue par l'autre, la période de S_2 sera deux fois celle de S_1, et par conséquent la fréquence de S_2 ne sera que la moitié de celle de S_1.

Si, au contraire, on commence par donner à S_1 et S_2 le même déplacement (fig.10) on sait que la diminution de volume de R_1 est deux fois celle de R_2. L'accroissement de pression dans R_1 est donc deux fois celui de R_2, c.-à-d. $2\Delta P$ contre ΔP. Ces valeurs sont substituées dans les expressions pour l'accélération:

[7.1] $G_1 = \frac{2\Delta P}{B \cdot \rho}$ [7.2] $G_2 = \frac{\Delta P}{B \cdot \rho}$ c.-à-d.

[7.3] $G_1 = 2G_2$

La vitesse moyenne de S_1 sera deux fois celle de S_2, et comme elles ont la même élongation, la fréquence de S_1 sera deux fois celle de S_2.

Puisque la vélocité n'est pas concentrée à l'ouverture, mais est distribuée dans toute l'étendue du tuyau (fig.2), on peut conclure qu'une fermeture aura le maximum d'effet sur la première résonance près de l'ouverture, où les variations de vélocité sont à leur maximum.

Si l'on regarde les seconde, troisième etc. résonances, il faut considérer de nouveau la distribution des variations de la vélocité dans toute l'étendue du tuyau. A la seconde résonance (fig.11) il y a deux endroits où la vélocité est maximale, à savoir à une distance de 1/3 Lcm de l'extrémité fermée et à l'ouverture du tuyau. A la troisième résonance il y aura trois endroits, à la quatrième quatre endroits, etc., où une fermeture aura un effet appréciable sur la fréquence. Chaque résonance a un maximum de vélocité à l'ouverture du tuyau et par suite chaque résonance baisse à cause d'une fermeture près de l'ouverture (mais pas toujours à un degré égal, voir le résumé ci-dessous).

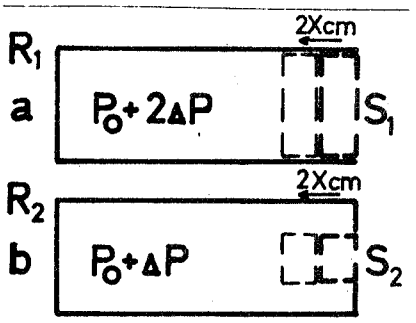


Figure 10

La relation entre l'élongation initiale de S_1 et S_2 et l'accroissement de pression dans le tuyau uniforme.

The relationship between initial elongation of S_1 and S_2 and pressure increment in the uniform pipe.

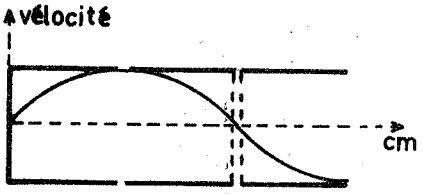


Figure 11

La distribution de vélocité dans un tuyau uniforme à sa seconde résonance.

The velocity distribution in a uniform pipe at its second resonance frequency.

RESUME

Une augmentation du volume du tuyau produit une diminution d'une résonance (et vice versa) d'autant plus qu'elle est plus proche d'un maximum de pression pour cette résonance, et une fermeture produit une diminution d'une résonance (et vice versa) d'autant plus qu'elle s'approche d'un maximum de vélocité pour

cette résonance, et, toutes choses égales par ailleurs, plus le changement du volume ou de l'ouverture est grand, plus le changement de la fréquence est important. Cependant, en pratique on ne peut pas séparer ces deux types de modification dans le conduit vocal. A cause des limitations imposées par les organes articulatoires, les variations du diamètre du conduit vocal entraînent simultanément des changements de volume et des fermetures/ouvertures. D'où, la formulation générale: Quand une partie d'un tuyau est rétrécie la fréquence de sa résonance devient grave ou aigue selon que la position de rétrécissement est près du point de vélocité maximum ou près du point de pression maximum.

Il en ressort que si la constriction (ou l'élargissement) du conduit vocal est située exactement entre un maximum de pression et un maximum de vélocité elle n'aura aucun effet. De plus: le conduit vocal est un système intégré dont la configuration est déterminée par la position de la langue et des lèvres. La langue ne peut pas accomplir simultanément des constrictions pharyngale et palatale étendues, au contraire, une constriction pharyngale produit une ouverture relative du conduit sous le palais dur, et vice versa, voir la fig.7. Le résultat cumulatif est une augmentation ([ɑ]) ou une diminution ([i]) "double" de la première résonance.

Ci-dessus on a considéré l'effet acoustique des changements de diamètre du conduit vocal pour chaque résonance en isolation. En pratique, les voyelles sont toujours composées de plusieurs résonances qui forment conjointement une oscillation complexe. Cela n'a aucune importance pour nos considérations: on peut traiter cette oscillation complexe comme une superposition d'oscillations sinusoïdales et observer l'effet des changements du conduit vocal pour chacun des composants.

Ce qui est plus important c'est qu'aussitôt qu'on ne prend pas comme point de départ un tuyau uniforme, mais un tuyau déja déformé (comme celui de la fig. 7 en bas, [i]) on ne peut pas quantifier les changements de fréquence de façon aussi simple qu'on a pu le faire ci-dessus pour les tuyaux uniformes. En effet la distribution des variations de pression et de vélocité n'est plus sinusoïdale (comme c'est le cas dans les fig.2, 8, 11). Un exemple suffira: pour le modèle de [i] la vélocité à la seconde résonance est presque nulle dans tout le tiers antérieur du tuyau, et une fermeture à l'ouverture du tuyau (arrondissement des lèvres) n'aura donc presque aucun effet sur la seconde résonance, mais comme la vélocité à la troisième résonance est maximale (plus grande, en effet, que pour le tuyau uniforme) elle diminue radicalement à la suite d'une fermeture aux lèvres. (Pour les diagrammes de la distribution de pression et de vélocité pour plusieurs voyelles, voir les trois références bibliographiques ci-dessous.)

CONCLUSION

Dans la parole les phénomènes sont beaucoup plus complexes que ne les décrit cet exposé: les parois du conduit vocal ne sont pas rigides, et il y a un rayonnement d'énergie appréciable vers l'extérieur. En plus de la perte d'énergie, le rayonnement occasionne un accordement des résonances qui est différent pour les fréquences hautes et basses, et qui est aussi fonction de l'ouverture aux lèvres. La fourniture sonore, c.-à-d. les impulsions de la glotte, constitue un élément qui complique encore la situation, entre autre choses à cause du couplage entre les cavités sub- et supraglottales. De plus, les éléments de physique et de mathématique utilisés ici ne suffisent pas en pratique - pour quantifier les conséquences acoustiques des mouvements de la langue et des lèvres, il faut faire le calcul des équations différentielles d'ordre supérieure.

REFERENCES

CHIBA, T., KAJIYAMA, M., 1958, The vowel, its nature and structure, Tokyo.
FANT, G., 1970, Acoustic theory of speech production, 2nd ed., Paris.
UNGEHEUER, G., 1962, Elemente einer akustischen Theorie der Vokalartikulation, Berlin.

THEME 3

ANALYSE DES INDICES ET DETECTION AUTOMATIQUE DES TRAITS

9èmes JOURNEES D'ETUDE SUR LA PAROLE

LANNION 31 mai - 2 juin 1978

CARACTERISTIQUES COMPAREES DE SYSTEMES RECONNAISSANT LA PAROLE

JEAN A. DREYFUS-GRAF, GENEVE (Suisse)
(5 Avenue de la Grenade)

RESUME

Toute communication implique un émetteur et un récepteur. Dans le cas de la reconnaissance automatique de la parole, l'émetteur est humain et le récepteur est un système artificiel. Sa modélisation, qui doit simuler les fonctions de l'oreille et du cerveau humains, dépend du but poursuivi (reconnaissance de la parole continue (phrases), découpée (mots) ou phonocodée), ainsi que du degré de complexité admis (nombres et qualités des locuteurs, des messages et des transmissions, téléphoniques ou non).

On compare les caractéristiques du système humain avec celles de divers systèmes artificiels, développés dans des laboratoires tels que : Université Carnégie-Mellon (HARPY), I.B.M., Bell-Laboratories, Threshold Technology, C.N.E.T. (CHARLES).

Les caractéristiques comparées concernent notamment les éléments suivants :

1. Modélisation des fonctions analytiques de l'oreille et algorithmiques du cerveau, 2. Hiérarchie des sources de connaissance ou contraintes :
   a) acoustiques (physiques), b) téléphoniques (avec et sans bruit),
   c) spectrales (vocoder), d) phonétiques (phonémiques et prosodiques),
   e) occurentielles (phonologiques), f) lexicales (vocabulaire), g) syntaxiques (grammaire), h) linguistiques complémentaires (sémantiques, contextuelles).
3. Moyens techniques et financiers mis en oeuvre, 4. Taux d'erreur obtenus aux divers niveaux des sources de connaissance, avec transmissions directes ou téléphoniques.

Autres sujets traités :

Analyse des anti-redondances inverses de systèmes symboliques, permettant d'évaluer l'importance relative des diverses sources de connaissance ou contraintes.

Perfectionnements aux phono-décodeurs (systèmes reconnaissant la parole codée, c'est-à-dire réduite à 6 ou 8 classes de phonèmes internationaux, par exemple). Les phono-décodeurs peuvent reconnaître des nombres non-limités de mots, prononcés par des locuteurs poly-idiomatiques, non-mémorisés par le système. Les taux d'erreur de la parole phonocodée (au niveau phonémique) doivent être 120 fois inférieurs à ceux de la parole naturelle, notamment dans les transmissions téléphoniques.

COMPARATIVE CHARACTERISTICS OF SPEECH-RECOGNIZING SYSTEMS

Jean A. DREYFUS-GRAF, GENEVA (Switzerland)
(5 Avenue de la Grenade)

SUMMARY:

All communication involves an emitter and a receiver. In the case of automatic speech recognition, the emitter is human and the receiver is an artificial system. Its modelling, which has to simulate the ear and brain functions, depends on the desired aim (recognition of continuous, segmented or phonocoded speech), as well as on the accepted complexity degree (number and quality of the speakers, messages and transmissions, whether direct or telephonic).

The characteristics of the human system are compared with those of various artificial systems developed in laboratories such as: Carnegie-Mellon University (HARPY), I.B.M., Bell Laboratories, Threshold Technology, C.N.E.T. (CHARLES).

The characteristics compared concern especially the following:

1. Modelling of the analytical ear functions and of the algorithmic brain functions.

2. Hierarchy of the knowledge sources or constraints: a) acoustic (physical), b) telephonic (with or without noise), c) spectral (vocoder), d) phonetic (phonemic and prosodic), e) occurrential (phonological), f) lexical (vocabulary), g) syntactical (grammar), h) linguistic complements (semantic, contextual).

3. Technical and financial resources.

4. Resulting error rates at the various levels of the knowledge sources, with direct or telephonic transmission.

Other subjects dealt with:

Inverse anti-redundancy analysis of symbolic systems, allowing an estimate of the relative importance of the various knowledge sources or constraints.

Improvements to phono-decoders (systems which recognize coded speech, for example, reduced to 6 or 8 international phoneme classes). Phono-decoders can recognize unlimited numbers of words, uttered by poly-idiomatic speakers, which are not memorized in the machine. The error rates of coded speech must be 120 times smaller (at the phonemic level) than those of the natural speech, especially over the telephone.

CARACTERISTIQUES COMPAREES DE SYSTEMES RECONNAISSANT LA PAROLE

Jean A. Dreyfus-Graf, Genève (Suisse)

1. Deux systèmes séparés par dix-huit années

Le champion actuel des systèmes reconnaissant la parole semble se nommer HARPY : il a été développé dans l'Université Carnegie-Mellon, l'une des bénéficiaires des 15 millions de dollars (LEA and SHOUP, 1977), distribués dans les années 1971 à 1976 par le Département Américain de la Défense (projet ARPA-SUR). Ce système reconnaît 1011 mots et 30 phrases à 3 mots, prononcés par 5 locuteurs (adaptés dynamiquement), avec des taux d'erreur de 3 et 5% (LOWERRE et REDDY 1977).

"HARPY ne parvient à ce résultat qu'en appliquant des restrictions sévères de grammaire et de prononciation, adaptées à chaque locuteur individuel... Les systèmes développés pendant le projet ARPA sont orientés vers des recherches de laboratoires et ne conviennent pas encore au monde réel." (WOLF, 1977).

Pourtant, dix-huit ans se sont déjà écoulés depuis que le Phonétographe III a reconnu les 31 phonèmes du français à l'intérieur d'un nombre illimité de mots, prononcés très lentement par un locuteur exercé. Ses taux d'erreur étaient de 4% sur les phonèmes et de 20% sur les mots (DREYFUS-GRAF, 1952, 1957, 1961, 1976, 1977). Ce premier "speech typewriter" est mentionné (FLANAGAN, 1976, réf.35) dans le récent numéro spécial "Man Machine Communication by Voice" des "Proceedings of the IEEE, April 1976".

2. Les degrés de complexité

Pour évaluer la qualité d'un système de reconnaissance, il ne suffit pas de distinguer 3 catégories de parole, telles que "phrases continues", "mots isolés", "syllabes phonocodées", mais il faut encore tenir compte d'autres facteurs, tels que : nombre et indépendance des locuteurs, transmission directe ou téléphonique, rapport signal/bruit, etc. Le tableau de la Fig. 1. résume 18 "degrés de complexité" (rangés par ordre décroissant), dont 3 pour les

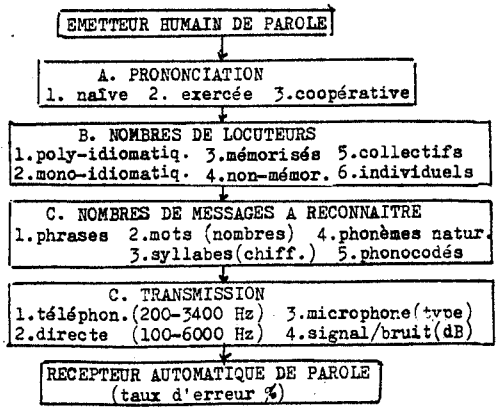


Fig.1 Degrés de complexité de la reconnaissance automatique de la parole (18 conditions)

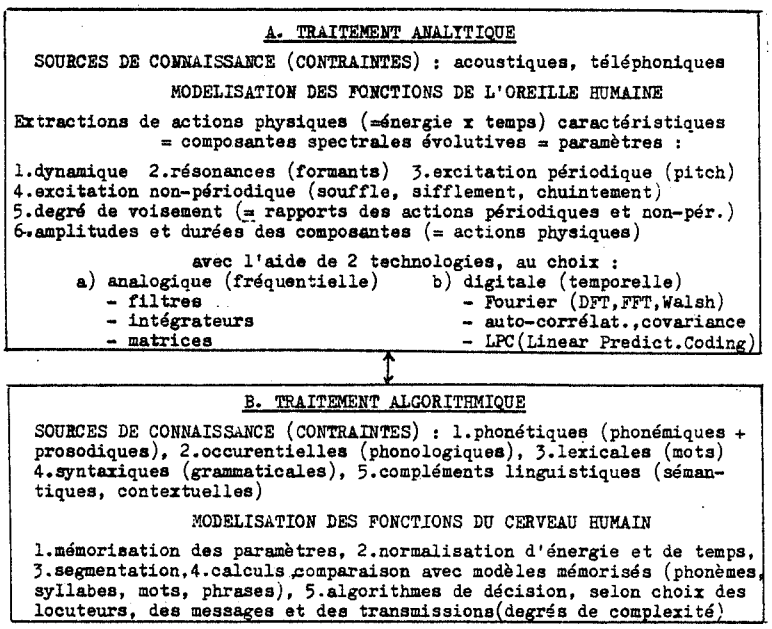


Fig.2 Modélisation du récepteur automatique de parole

prononciations, 6 pour les locuteurs, 5 pour les messages et 4 pour les transmissions.

Le choix de ces degrés dépend du but poursuivi; l'objectif le plus modeste est la reconnaissance directe de 10 chiffres articulés par un seul locuteur coopératif; le plus ambitieux est la reconnaissance téléphonique d'un nombre illimité de phrases, de mots et de nombres, prononcés par un nombre quelconque de locuteurs naïfs, dans une ambiance bruyante. Entre ces 2 extrèmes se trouvent toutes les gammes des recherches intermédiaires.

3. Modélisation du récepteur de parole

Toute communication implique un émetteur et un récepteur. Dans le cas de la reconnaissance automatique de la parole, l'émetteur est humain et le récepteur est une machine. C'est donc le récepteur de parole qu'il s'agit de modéliser. (La modélisation de l'émetteur (conduit vocal) est fondamentale dans le cas inverse du synthétiseur de parole ou vocoder).

Le tableau de la Fig.2. résume la modélisation du récepteur automatique de parole. Elle comprend une partie analytique et une partie algorithmique, simulant respectivement les fonctions de l'oreille et du cerveau humains.

Selon ses buts et moyens, chaque laboratoire de recherches expérimente sa propre modélisation, donnant la priorité tantôt à l'oreille, tantôt au cerveau.

Parmi les laboratoires les plus actifs, on peut citer les suivants, dont les publications sont précisées dans la bibliographie annexée :

1. Carnégie-Mellon University, Pittsburg, USA (LOWERRE, 1977)
2. Sperry Univac, St.-Paul, USA (MEDRESS et Al. 1977)
3. M.I.T.-Lincoln Lab., Lexington, USA (JOHNSON et WEINSTEIN, 1977; GOLD 1977)
4. Bolt, Beranek and Newman, Cambridge, USA (WOLF, 1977; COOK et SCHWARZ, 1977)
5. Bell Lab., Murray Hill, USA (RABINER and SAMBUR, 1976)
6. C.N.E.T., Lannion, France (EL MALLAWANY, 1976; GAGNOULET, MERCIER, VIVES et VAISSIERE, 1977; QUINTON 1977; COURBON, CARTIER et LORAND, 1977)
7. C.N.R.S., Orsay, France (MARIANI et LIENARD, 1978)
8. Université de Nancy I, France (HATON, MARI et PIERREL, 1977)
9. Techn. Universität, München, R.F.A. (HESS, 1976; RUSKE et SCHOTOLA, 1977)
10. Inst.f. Information Transmission, Moscou, URSS (TSEMEL, 1975)
11. I.B.M., Yorktown-Heights, USA (JELINEK, 1976, DIXON, 1977)
12. Threshold Technology, Delran, USA (HERSCHER, 1977)

4. Caractéristiques comparées de 5 systèmes techniques, avec référentiel humain

Parmi les systèmes développés dans les 12 laboratoires mentionnés ci-dessus on peut en sélectionner 5, particulièrement typiques : 1. SYSTEME CARNEGIE-MELLON (phrases et mots, avec plusieurs locuteurs individuels), 2. SYSTEME IBM (phrases et mots, avec locuteur unique), 3. SYSTEME BELL-LAB. (nombres, avec locuteurs collectifs). 4. SYSTEME THRESHOLD (mots et nombres, avec locuteurs individuels), 5. SYSTEME CHARLES du CNET (syllabes phonocodées avec locuteurs collectifs).

Le tableau de la Fig.3. en compare les caractéristiques principales, avec modélisations et taux d'erreur, dans 2 cas de transmission : directe et téléphonique.

"Modélisation" implique "modèle". Or celui-ci n'est autre que le récepteur de parole humain, élaboré par la nature au cours des 200 millions d'années qui se sont écoulées sur terre depuis l'apparition des premiers vertébrés, et qui obéit au principe cybernétique universel "Maximum de débit d'information (bit/sec) compatible avec un minimum d'action (watt.sec^2)". (DREYFUS-GRAF, 1970, 1972, 1974, 1976).

Ainsi les taux d'intelligibilité (ou d'erreur) du "système humain" figureront dans la Fig. 3. comme référentiel.

I. LE SYSTEME HUMAIN (référentiel)

a) L'intelligibilité aux syllabes dépourvues de sens (netteté aux logatomes) ne fait appel qu'aux niveaux inférieurs des sources de connaissance, tels qu'acoustiques (ou téléphoniques), phonétiques et occurentiels, (phonologiques).

Les taux d'erreur en sont environ : 2% en direct, 12% par téléphone (CHAVASSE, 1951; LORAND, 1969), dans le cas d'une langue naturelle, telle que le français, exigeant quelque 32 phonèmes. Par contre, dans le cas d'une langue simplifiée (phonocode ŜOKINA), se contenant de 6 phonèmes, les taux d'erreur sont 10 fois meilleurs en direct, et 120 fois meilleurs par téléphone. (DREYFUS-GRAF, 1973, 1975, 1976; KÖSTER 1976, PERSON, 1974)

Ces taux d'erreur restent constants, tant que le rapport signal/bruit n'est pas inférieur à 20 dB dans le cas du français, respectivement à 6 dB dans le cas du phonocode ŜOKINA.

DIR= direct 100-6000 Hz
TEL=téléphon 200-3400 Hz

SOURCES DE CONNAISSANCE OU CONTRAINTES (niveaux sup.) P= phonétique O= occurentiel L= lexical S= syntaxique C= compléments linguistiques		TAUX D'ERREUR (%) DANS LA RECONNAISSANCE DE LA PAROLE													
		PAR L'HOMME				PAR LA MACHINE, AVEC LE SYSTEME No :									
		I. intelligibilité avec opérateurs exercés (nombres non limités)				II. CARN.MELL. UNIVERSITY (Lowerre, Reddy,)		III. I.B.M. (Jelinek Dixon, & Al.)		IV. BELL-LAB (Rabiner, Sambur, & Al.)		V. THRESHOLD TECHNOLOGY (Martin, Herscher)		VI. C.N.E.T, CHARLES (Courbon, Cartier, Lorand,)	
		langue naturelle 32 phonèmes 25000 mots 10^{480} phrases		langue phonocodée ŜOKINA 6 phonèm. 25000 mots		30 phrases à 3 mots, parmi 1011 5 locuteurs adaptés		363 phrases, 250 mots 1 locuteur individ.		20 mots (nombres à 3chiff) 10 locuteurs collectifs		75 mots (nombres à 4chiff) 30 locuteurs collect. (M+F)		phonocode SOTINA 81 mots CV+CVCV 15 locut. collect. (M+F)	
		DIR	TEL	DIR	TEL	DIR	TEL	DIR	TEL	DIR	TEL	DIR	TEL	DIR	TEL
		%	%	%	%	%	%	%	%	%	%	%	%	%	%
P	PHONEMES	0,7	4,2	0,07	0,035	46	-	38	-		-	8			3
O	SYLLABES (phonatomes, chiffres à 3 phonèmes)	2	12	0,2	0,1	-	-	-	-	-	-	2,1	3,2	-	-
	MOTS (nombres à 3,4 chiffres)	6	30	0,6	0,3	-	-	-	-	-	-	-	-	-	1,5
L		-	-	-	-	3	-	3	-	9	36	7,5	12	-	-
S	PHRASES	0	2	-	-	5	11	19	-	-	-	-	-	-	-
C						-	-	-	-	-	-	-	-	-	-
CARACTERISTIQUES DES SYSTEMES ARTIFICIELS	PARTIE ANALYTIQUE (modélisation des fonctions de l'oreille)					10kHz, 9b. autoc. LPC parcor 64 pitch 6 ampl. bit 5		20kHz, 15b FFT, LPC 1117/62 "phones"		ZCR log.énerg LPC2pôles coeff. autocorr.		16 canaux 1/3 oct. E, dE/df 512 b/s		6 canaux +plosiv. lin/log comb.lin.	
	extracteur de pitch					oui		oui		non		non		non	
	PARTIE ALGORITHMIQUE (modélisation des fonctions du cerveau)					heuristiq. 98phones 1265bit/s adapt.dyn. au locut.		stochast. (Viterbi+ Markow) temps normalisé		arbre de décision probabiliste		temps normalisé 16 class. dist.min.		6 class. phonèm. +contrainte CV, CVCV	
	CALCULATEUR type					PDP 10		360/91+ 370/168		-		LSI 11		IMP 16 c	
	MEMOIRE (kbytes)					256 x 36				-		16384 bit / 32 mots		1+0,5 ROM+RAM	
	INSTRUCTIONS/sec de parole (mio)					17-30		30				0,5		0,1	
	facteur de temps réel					10 à 50x		30 x		-		1 x		1 x	
	FINANCEMENT					part. de ARPA-SUS		-		-		TTI+Gouvern.USA		CNET+ CRI+IRIA	
	CREDIT R & D					$ 15 mio		-		-		$ 6 mio		FF 0,5mio	
	ANNEES					1971-76		-		-		1960-77		1973-76	

Fig.3 Caractéristiques comparées de la reconnaissance de la parole, par le système humain (référentiel I.) et par 5 systèmes artificiels (II, III, IV, V, VI)

Il peut paraître paradoxal que l'intelligibilité des syllabes phonocodées est 2 fois meilleure par téléphone qu'en direct, alors que celle des syllabes naturelles est 6 fois moins bonne. L'explication, c'est que le filtre 200-3400 Hz dégrade la plupart des 32 phonèmes d'une langue naturelle, ainsi qu'une partie des bruits, tandis qu'elle respecte les phonèmes O, I, A, Ŝ(ch), K, N du phonocode ŜOKINA.

Ces considérations préliminaires permettent déjà de formuler deux conclusions fondamentales :

1. En transmission directe (non téléphonique), le récepteur humain opère la presque totalité de la reconnaissance de la parole (98%), sans avoir besoin de faire appel aux niveaux supérieurs des sources de connaissance linguistiques, telles que lexicales, syntaxiques ou sémantiques. En effet, celles-ci n'apportent que 2% à la compréhension complète.

2. Dans le cas de la transmission téléphonique, une langue simplifiée, telle que basée sur le phonocode ŜOKINA, présente des taux d'erreur (0,1%) qui sont 120 fois meilleurs que ceux d'une langue naturelle, telle que le français (12%). Cette supériorité augmente encore, au fur et à mesure que le rapport signal/bruit se détériore, jusqu'à une limite inférieure de 6 dB.

b) L'intelligibilité (ou netteté) aux phonèmes N_p peut être déduite de l'intelligibilité aux syllabes $N_s = 1 - E_s$ (qui est le complément à 1 de l'erreur E_s) grâce à la formule $N_p = (N_s$ élevé à la puissance $1/P)$, où P = nombre de phonèmes (moyen) par syllabe. Ainsi, avec P=3 phonèmes par syllabe, les taux d'erreur aux phonèmes $E_p = 1 - N_p$ deviennent :

en transmission directe ..... 0,7% pour le français et 0,07% pour ŜOKINA
" " téléphonique ..... 4,2% " " " " 0,035% " "

Ces excellentes performances phonémiques de base permettent ensuite au récepteur humain de reconnaître, sans erreur, des messages linguistiques sélectionnés parmi 25.000 mots et 10^{480} phrases, par exemple.

II. LE SYSTEME CARNEGIE-MELLON (HARPY) est présenté dans les colonnes correspondantes de la Fig.3. avec ses degrés de complexité et éléments de modélisation.

En transmission directe, ce système reconnaît 1011 mots et 30 phrases à 3 mots, prononcés par 5 locuteurs adaptés individuellement avec des taux d'erreur de 3% sur les mots et de 5% sur les phrases (LOWERRE et REDDY, 1977; LOWERRE 1977; WOLF, 1977). En transmission téléphonique, ce taux passe à 11%.

Cependant, au niveau de base phonémique, l'erreur présente un taux très élevé de 46%, c'est-à-dire 65 fois plus grand que celui du récepteur de parole humain (0,7%). Dans un chapitre ultérieur, nous aurons recours à l'analyse des anti-redondances inverses pour éclairer ces divergences.

III. LE SYSTEME IBM. (JELINEK 1976; DIXON 1977) suscite des remarques similaires. Il reconnaît 250 mots et 363 phrases, prononcés par un locuteur adapté, avec des taux d'erreur respectifs de 3% et 19%. Et le taux d'erreur au niveau phonémique est encore de 38%.

IV. LE SYSTEME BELL-LAB. reconnaît 20 nombres à 3 chiffres, prononcés par 10 locuteurs collectifs, avec des taux d'erreur de 9% ou de 36%, selon que la transmission est directe ou téléphonique. (RABINER et SAMBUR, 1976). Le taux d'erreur au niveau phonémique n'est pas communiqué, mais semble très élevé, comme pour les 2 systèmes précédents.

V. LE SYSTEME THRESHOLD (VIP 100 ou 500) reconnaît, par exemple, 75 nombres à 4 chiffres, prononcés par 30 locuteurs (mémorisés, coopératifs et individuels) avec des taux d'erreur de 2,1% aux chiffres et de 7,5% aux nombres. Le taux d'erreur aux phonèmes est de 8%, c'est-à-dire qu'il est très inférieur à ceux des systèmes "Carnégie-Mellon" (46%) et "I.B.M." (38%). Le système "Threshold" est d'ailleurs le seul qui soit utilisable pratiquement, et qu'on trouve sur le marché (25.000.-$). L'analyse s'effectue avec 16 filtres "tiers d'octave" (HERSCHER, 1977).

VI. LE SYSTEME CHARLES du CNET ne reconnaît que des mots à 1 ou 2 syllabes (CV ou CVCV) formés à partir de 5 ou 6 phonèmes du phonocode ŜOKINA. Celui-ci est limité à 3 classes de voyelles O,I,A, et à 3 classes de consonnes Ŝ(ch), K, N. Il a reconnu, par exemple, 81 mots (dont 6 mots CV et 75 mots CVCV) prononcés par 15 locuteurs exercés, mais non-mémorisés et collectifs (masculins et féminins), avec un taux d'erreur de 1,5%. Le taux d'erreur au phonème était de 3%. Ainsi, le taux d'erreur aux mots à 2 syllabes (= 4 phonèmes) serait de 12% (3 x 4), si la contrainte structurelle CVCV (consonne-voyelle-consonne-voyelle) n'intervenait pas pour le réduire aux 1,5% obtenus. (COURBON, CARTIER et LORAND, 1977).

Le taux d'erreur de 3% sur les phonèmes ne correspond pas à l'amélioration attendue, quand on passe de la parole naturelle, avec 32 phonèmes, à la parole codée avec 6 phonèmes, et qui devrait présenter un facteur 10 en direct, respectivement 120 par téléphone. Il ne permet pas d'atteindre des objectifs assignés aux phonocodes, tels que ŜOKINA (DREYFUS-GRAF, 1972, 1974 1975, 1976, 1977), notamment la reconnaissance des nombres (phonocodés) entre zéro et l'infini, prononcés téléphoniquement par des locuteurs poly-idiomatiques (de toutes nationalités).

Les chapitres suivants esquisseront des méthodes qui ont pour but d'améliorer la modélisation de systèmes reconnaissant la parole, tant naturelle que codée.

Fig.4 FORMULES POUR L'ANALYSE DES REDONDANCES DE SYSTEMES SYMBOLIQUES

a) Débit d'information spectral D_d d'un canal (largeur de bande L(Hz)), transmettant d degrés d'ammplitude à la fréquence $F_d = 2.L(Hz)$:

(1) $\boxed{D_d(bit/sec) = R_d(bit).F_d(Hz)}$ où $R_d = \log_2 d$ (bit), ou dynamique (dB)

b) Débit d'information symbolique D_s d'un canal transmettant s symboles équiprobables ($p_i = 1/s$) à la fréquence F_s (symb/sec), avec $N_s = \log_2 s$ (bit/symb) = entropie négative

(2) $\boxed{D_s(bit/sec) = N_s(bit/symb).F_s(symb/sec)}$

Si les s symboles ne sont pas équiprobables, avec $p_i \neq 1/s$:

(3) $\boxed{N = -\sum p_i.\log_2 p_i \ (bit/symb)}$ = entropie négative moyenne = néguentropie

c) Redondance R_{xy}, quand le débit d'information $D_x = N_x.F_x$ (bit/sec) passe à un autre débit d'information (plus petit) $D_y = N_y.F_y$(bit/sec), avec $F_x = F_y$ et $N_x \geqslant N_y$

(4) $\boxed{R_{xy} = 1 - N_y/N_x \ (bit/bit)}$, où N_y/N_x = rapport des néguentropies = anti-redondance = $A_{xy} = 1 - R_{xy}$

d) Anti-redondance (= 1 - redondance)

(5a) $\boxed{A_{xy} = 1 - R_{xy} = N_y/N_x \ (bit/bit)}$

Anti-redondance inverse (bit/bit)

(5b) $\boxed{A_{yx} = 1/A_{xy} = N_x/N_y = N_x/N_{x+1}}$

e) Redondances $R_1, R_2, \ldots R_n$ en série = S_n

La résultante S_n de n redondances $(R_1, R_2, \ldots R_n)$ en série correspond au produit des n anti-redondances $(A_1, A_2, \ldots A_n)$

(6) $\boxed{S_n = 1 - (A_1.A_2 \ldots A_n)}$

f) Redondances $R_1, R_2, \ldots R_n$ en parallèle = P_n

La résultante P_n de n redondances $(R_1, R_2, \ldots R_n)$ en parallèle correspond à la somme des n anti-redondances $(A_1, A_2, \ldots A_n)$

(7) $\boxed{P_n = 1 - (A_1 + A_2 + \ldots + A_n)}$

5. Sources de connaissance et anti-redondances inverses

Les systèmes comparés précédemment puisaient de manière empirique leurs informations aux diverses sources de connaissance; plus les taux d'erreur étaient élevés au niveau phonémique, et plus les algorithmes correctifs prenaient de l'importance aux niveaux supérieurs, tels que syntaxiques.

Ne peut-on trouver une méthode qui permette d'évaluer d'emblée les importances relatives des sources de connaissance ? "L'analyse des redondances de système symboliques" (DREYFUS-GRAF, 1976, 1977) a montré un premier fil directeur à ce sujet. Nous proposons maintenant "l'analyse des anti-redondances inverses" qui s'adapte mieux à la diversité des buts et degrés de complexité.

Le tableau de la Fig. 4. rappelle les formules de la redondance R et de l'anti-redondance A = 1 - R. L'anti-redondance inverse est 1/A, selon la formule (5b). Ainsi quand un système passe d'une néguentropie N_x à une autre $N_y = N_{x+1}$ (plus petite), l'antiredondance inverse est $A_{x+1} = N_x / N_{x+1}$ et elle caractérise l'importance relative (IRSC %) de la source de connaissance ou contrainte x+1.

Le tableau de la Fig.5. applique l'analyse des anti-redondances inverses, $A_{x+1} = N_x / N_{x+1}$, pour évaluer l'importance relative (IRSC %) des 8 principales sources de connaissance ou contraintes, à savoir :

1. acoustiques, 2. téléphoniques, 3. spectrales (vocoder), 4. phonétiques (= phonémiques + prosodiques), 5. occurentielles (phonologiques), 6. lexicales (vocabulaire), 7. syntaxiques (grammaticales), 8. linguistiques complémentaires (sémantiques + contextuelles).

TAUX D'IMPORTANCE (%) DES SOURCES DE CONNAISSANCE ou CONTRAINTES, en fonction des ANTIREDONDANCES INVERSES $A_{x+1} = N_x / N_{x+1}$
$N_x = (D_d / D_s)_x$ = néguentropie (bit/symbole)
D_d = débit d'information (bit/sec)
D_s = débit de symboles (symboles/sec)

x	indice d'analyse	SOURCE DE CONNAISSANCE avec débit d'information bit/sec	ELEMENT RECONNU	niveau de reconnais	1. PAROLE CONTINUE 32 phonèmes 25.000 mots symb=phonèm. 10 symb/sec			2. PAROLE DECOUPEE 32 phonèmes 1000 mots symbole=mot 2 symb/sec			3. PAROLE PHONOCODEE 6 phonèmes 25.000 mots symb=phonem. 10 symb/sec		
					N_x	N_x/N_{x+1} A_{x+1}	%	N_x	N_x/N_{x+1} A_{x+1}	%	N_x	N_x/N_{x+1} A_{x+1}	%
1	a	acoustique 144.000 bit/s	-	-	14400	---	---	72000	---	---	14400	---	---
						2,9	6		2,9	2		2,9	3
2	t	téléphonique 50.000 bit/s	-	-	5000	---	---	25000	---	---	5000	---	---
						20,8	40		20,8	15		62,4	64
3	v	spectrale (vocoder) 2.400 bit/s	-	-	240	---	---	1200	---	---	80	---	---
						20	39					31	31,0
4	p	phonétique= phonèm.+pros. 120 bit/s	phonèmes +signes auxiliair.	P	12= 6+6	2,9	6		120	83	2,6	---	---
5	c1	occurentiel (phonologiq.) 42 bit/s	syllabes logatomes chiffres	O	4,2	---	---	10					
						1,4	2					1,3	1,4
6	c2	lexicale (vocabulaire) 30 bit/s	mots nombres	L	3	---	---		---	---	2		
						2	4		-	-			
7	c3	syntaxique (grammaire) 15 bit/s	phrases	S	1,5	---	---	-	---	---		---	---
						1,5	3		-	-		-	-
8	c4 c5	compléments linguistiques 10 bit/s	textes	C	1	---	---	-	---	---	-	---	---
		total				51,5	100		143,7	100		97,7	100

Fig.5 Taux d'importance relative (%) des sources de connaissance ou contraintes, en fonction des anti-redondances inverses

L'analyse est effectuée dans 3 cas typiques de la reconnaissance de la parole :

a) parole continue (phrases de langue naturelle, 25.000 mots usuels, 32 phonèmes)
b) parole découpée (1000 mots isolés de la langue naturelle)
c) parole phonocodée (25.000 mots ou nombres, phonocode ŜOKINA, 6 phonèmes)

Dans les cas a) et c), 85% (=6+40+39), respectivement 98,7% (= 3+64+31,7) de la reconnaissance devrait être achevés au niveau No 4, phonétique (= phonémique + prosodique). Le tableau de la Fig. 3. montre que les taux d'erreur relatifs des systèmes II, III et IV sont très éloignés de ces pourcentages, tandis que leurs vocabulaires sont très inférieurs aux 25.000 mots usuels d'une langue naturelle.

Dans le cas b), 17% de la reconnaissance doit être achevés au niveau No 3, spectral, puis 83% entre ce niveau No 3 et le niveau No 6, lexical. En effet, dans ce cas, le symbole élémentaire n'est plus un phonème, avec un débit symbolique de 10 phonèmes par seconde, mais un mot (à 5 phonèmes, par exemple) avec un débit symbolique de 2 mots par seconde.

Une étude ultérieure développera d'une manière plus détaillée l'analyse des anti-redondances inverses, en vue de modéliser divers types de récepteurs de parole en fonction des buts poursuivis et des degrés de complexité acceptés.

Pour l'instant, nous nous bornerons à proposer des perfectionnements aux systèmes reconnaissant la parole codée, tels que CHARLES, et qu'on peut nommer "phonodécodeurs". Théoriquement, la parole phonocodée devrait être 120 fois supérieure à la parole naturelle, notamment quand il s'agit de faire reconnaître de grands nombres de chiffres à travers le téléphone.

6. Perfectionnements proposés pour les phono-décodeurs

Nous proposons le développement d'un prototype industriel de phono-décodeur ŜOKINARE, reconnaissant jusqu'à 4 classes de voyelles, O,A,I,E, et 4 classes de consonnes, Ŝ(=ch), K, N, R (roulé). Il inclut toutes les combinaisons partielles, aisément prononçables, de ces 8 classes de phonèmes, donc aussi les phonocodes ŜOKINA, ŜOKEMA, ŜOKI, ŜOKA, etc. Nous en avons développé le nouveau schéma, réalisable en technologie digitale, et qui élimine les défauts du premier prototype. Il admet des taux d'erreur inférieurs à 0,2% au niveau phonémique. Il sépare les voyelles et les consommes sélectionnées, des autres sons. Il permet la reconnaissance de mots en nombres non-limités, et notamment les nombres phonocodés, prononcés par des locuteurs poly-idiomatiques et non-mémorisés.

Un tel phono-décodeur ŜOKINARE serait susceptible d'applications immédiates et généralisées, contrairement aux systèmes reconnaissant la parole naturelle, dont les vocabulaires sont restreints, qui doivent s'adapter individuellement à chaque locuteur, et dont les liaisons téléphoniques augmentent les taux d'erreur. Son prix pourrait se situer entre celui d'un micro-ordinateur et celui d'un calculateur de poche, selon l'importance de son vocabulaire.

Bibliographie

systèmes artificiels

COOK, C.C., SCHWARZ, R.M., Advanced Acoustic Techniques in Automatic Speech Understanding, IEEE-ICASSP, Hartford, 1977

COURBON, J.L., CARTIER, M., LORAND, P., Reconnaissance automatique de la parole "CHARLES" Revue d'Acoustique, Paris, Vol.10-No.43- 1977

DIXON, N.R., SILVERMAN, H.F., The 1976 Modular Acoustic Processor (MAP). IEEE-ASSP. Vol.25, No 5, October 1977, pp 367-379

DREYFUS-GRAF, J.A.,

a) Le typo-sonographe phonétique ou phonétographe (Phonétographe I) Bull. Techn. PTT, Berne, No 12/1952, pp 363-379
b) Phonétographe et sub-formants (Phonétographe II). Bull.Techn.PTT, Berne, No 2/1957, pp 41-59
c) Phonétographe : Présent et futur (Phonétographe III). Bull. Techn. PTT, Berne, No 5/1961, pp 160-172.
d) La parole humaine et l'informatique (Phonétographe V, vocographe phonacteur, mélographe, etc.) Automatisme, Dunod,Paris,No.9,1970
e) Physique et Cybernétique (Le monde physique comme système de signaux auto-régulés), Helv.Phys.Acta, Vol.45, 1972,pp.966-970
f) Tests d'intelligibilité de la parole codée (Phonocodes) Symposium GALF, FASE, APAV, Liège 1973 (Rapports)
g) "Codes phonétiques (Phonocodes) et télécommunications" FASE 75, Paris, Colloque No. 1, 1975
h) "Machines commandées par la parole humaine, naturelle ou codée" Vol.ISR 32, pp351-360, Birkhäuser-Verlag, Bâle, 1976
i) "Redundancy Analysis : A Guide Line through Speech Recognition" 1977-IEEE-ICASSP, Hartford (Conn.), pp. 823-826

EL MALLAWANY, I., 1976, Le dialogue homme-machine : le traitement du signal, Annales des Télécommunications, tome 31, No 11-12, 1976

FLANAGAN, J.,L., Computers that Talk and Listen : Man Machine Communication by Voice, Proceedings of the IEEE, April 1976

GAGNOULET, C., MERCIER, G., VIVES, R., VAISSIERE, J., A Multi Purpose Speech Understanding System, IEEE-ICASSP, Hartford, 1977

GOLD, B., Digital Speech Networks, Proc. IEEE, Dec. 1977

HATON, J.P., MARI, J.P., PIERREL, J.M., Research Towards Speech Understanding Models for Artificial and Natural Languages,IEEE-ICASSP,Hartford,1977

HERSCHER, M.B., Real-Time Interactive Speech Technology at Threshold Technology NASA, Ames, Dec. 1977

HESS, W.J., ...Phonemic Recognition of Speech. IEEE, ASP, Febr.1976, pp 14-25

JELINEK,F.,Continous Speech Recognition by Statistical Methods,Proc.IEEE,Apr.76

JOHNSON, D.H., WEINSTEIN, C.J., A Real Time Phrase Recognizer Using Telephon Bandwidth Speech, IEEE, ICASSP, Hartford, 1977

LEA, W.A., SHOUP, J.E., Specific Contributions of the ARPA SUR Project to Speech Science, Meeting of the ASA, Miami, Dec. 1977.

LOWERRE,B.T.,REDDY, D.R.,Summary of Results, Carnegie-Mellon University,Aug.1977

LOWERRE, B.T., Dynamic Speech Adaptation in the Harpy Speech Recognition System, IEEE-ICASSP, Hartford 1977

MARIANI, J.J., LIENARD, J.S., ESOPE, Congrès AFCET-IRIA, févr. 1978

MEDRESS, M.F., et Al., A System for the Recognition of Spoken Connected Word Sequences, IEEE-ICASSP, Hartford, 1977

RABINER, L.R., SAMBUR, M.R., Some Preliminary Experiments in the Recognition of Connected Digits, IEEE-ASSP, April 1976

RUSKE, G., SCHOTOLA, T., An Automatic Speech Recognition System Using Syllable Nuclei and Consonant Clusters, 9.ICA, Madrid, 1977

QUINTON, P., Utilisation d'un analyseur syntaxique pour la reconnaissance de la parole continue, Annales des Télécommunications, No 9-10, 1977

TSEMEL, G., "Rechevoye obshcheniye v avtomatizierovannykh systemakh", Nauka, 1975, pp 5-33

WOLF, J.J., HWIM, A Natural Language Speech Understander, IEEE Conference on Decision and Control. New Orleans. Dec. 1977

système humain (intelligibilité)

CHAVASSE, P., L'application des moyens d'analyse de la qualité des transmissions téléphoniques, La Cybernétique, Paris 1951

KÖSTER, J.-P., Phonokodes und die Perzeption Konstruierter Sprachen, Hamburger Phonetische Beiträge, 17 (1976), pp 35-82 (phonocode SOTINAKEMUS)

LORAND, P., Influence d'un bruit blanc sur la netteté d'une communication téléphonique, C.N.E.T., Lannion, Note Technique ETA/1,(1969)

PERSON, J.-M., et Coll., Essais subjectifs de netteté aux "Sonatomes" (SOTINAKEMUS), C.N.E.T., Lannion, Fiche Téléphonométrique No 2038 (1974)

9èmes JOURNEES D'ETUDE SUR LA PAROLE

LANNION 31 mai - 2 juin 1978

LOCALISATION DES VOYELLES DANS LE PLAN (F_1,F_2)

APPLICATION A LA RECONNAISSANCE DE LA PAROLE

Guy PERENNOU et Jean CAELEN

Laboratoire C.E.R.F.I.A.

118, route de Narbonne - 31077 TOULOUSE

.**RESUME**

Le modèle mathématique d'oreille mis au point par J. CAELEN permet une analyse spectrale dégageant nettement la structure formantique. En partant d'une telle analyse du signal, on étudie la répartition des voyelles dans le plan F_1- F_2pour un locuteur donné ainsi que l'influence de quelques consonnes sur cette répartition.

On montre comment des ambiguités peuvent être levées par l'utilisation de F_3 et de F_4 et l'on aborde également le cas des voyelles compactes.

Malgré la bonne qualité de la reconnaissance des phonèmes, il subsiste toujours un taux plus ou moins important d'incertitudes. Certaines d'entre elles peuvent se résoudre au niveau de la syllabe.

Dans ce but, on expose dans ses grandes lignes une méthode basée sur l'influence spécifique des phonèmes sur une voyelle adjacente.

Cela permet d'éliminer des associations où s'expriment des distortions contraires aux influences qu'elles devraient avoir mutuellement.

SUMMARY

VOWELS LOCALIZATION IN THE PLANE (F_1, F_2)

APPLICATION TO SPEECH RECOGNITION

Guy PERENNOU et Jean CAELEN
Laboratoire C.E.R.F.I.A.
118, route de Narbonne- 31077 TOULOUSE

The mathematical model of ear perfected by CAELEN J. is allowing spectral analysis which makes evident formantic structure.

Taking into account such an analysis of signal, we may study vowel distribution in the plane(F_1, F_2) for a given locutor then, the effect of a few consonants upon this distribution is observed.

We explain how to remove ambiguities by using F_3 and F_4 and the case of compact vowels is approached.

In spite of the good quality of phonemic recognition, there always stands a rate of uncertainty more or less important. Some uncertainties may be resolved by considering syllabic constraints.

To this end, we outline a method based on specific influence of phonemes upon an adjacent vowel. In this way, we can suppress associations where appear distortions opposite to the influences which they should have reciprocally.

INTRODUCTION

Des expériences perceptives ont montré que l'identification des voyelles s'effectue essentiellement à l'aide des formants F_1 et F_2 (voir par exemple "BECKMANS, R.,1977).

L'espace perceptif serait ainsi constitué d'un triangle vocalique, s'inscrivant dans le plan (F_1 , F_2), sur lequel chaque voyelle se localiserait.

L'adaptation au locuteur se traduirait par la déformation du triangle vocalique à partir d'informations de deux types : le fondamental et la répartition d'ensemble des phonèmes entendus (ce dernier point restant encore à approfondir).

Nos études préliminaires à la reconnaissance automatique des voyelles confirment la prépondérance du plan (F_1 , F_2). Cependant, il est possible d'adjoindre certains indices améliorant les performances. Ainsi pour la discrimination entre /i/ et /y/ il est bon d'utiliser F_3 (PERENNOU,DOURS,FACCA, 1975) (LANDERCY, A. et RENARD, R., 1975). Du reste, pour leurs synthèses, on relève souvent la fréquence de F_2 pour pallier l'absence de F_3 . Notons également l'essai de structuration d'un espace à trois dimensions F_1 , F_2 , F_3 proposé par (D.J. BROAD et HISASHI WAKITO, 1977).

L'objet de notre article est de montrer comment, en liaison avec d'autres indices (F_3, compacité) l'information contenue dans le couple (F_1 , F_2) peut être utilisée à la fois de manière statique et de manière dynamique (effet de coarticulation). Nous n'entendons nullement produire et fonder des indices nouveaux mais plutôt introduire certaines données connues dans le mécanisme de reconnaissance au niveau phonétique.

Notons à cet égard

1) que le processus de reconnaissance chez l'homme intègre les indices présents dans le signal, La présence de tous ceux qui participent à un trait n'est pas obligatoire pour sa perception ;

2) que dans les expériences perceptives les indices sont,autant que possible,étudiés isolément pour établir leur contribution individuelle ;

3) dans les processus de reconnaissance l'étiquetage en classe peut se faire à partir d'ensembles de paramètres sans que leur pertinence soit démontrée : certaines particularités du signal peuvent accompagner tel phonème de manière fréquente sans que leur participation à la perception soit positive ; cela ne va pas sans poser des problèmes lorsque l'on change de locuteur.

DONNEES ET ANALYSE SPECTRALE

Les données sont constituées par une série d'enregistrements de 19 phrases pour un locuteur homme. Le signal a été échantillonné à 15000 Hz et analysé par un système utilisant le modèle d'oreille de (J. CAELEN, 1974) (on peut également consulter : J.CAELEN, M.C. EL JAI, G. PERENNOU, 1977). Des enregistrements provenant de quatre autres locuteurs sont également pris en considération pour la confirmation des faits observés. A cet égard, précisons que les fréquences de crête dans le spectre cochléaire redonnent les fréquences sous-jacentes du signal, sauf si elles sont trop rapprochées. Pour fixer les idées, à amplitude égale, l'effet d'attraction mutuelle des deux maxima apparaît à partir d'un rapport d'un tiers d'octave.

L'intérêt de l'analyse basée sur ce modèle d'oreille est qu'elle donne des spectres d'où se dégage clairement la structure formantique aussi bien dans les transitions que dans les parties stables.

Les paramètres relevés sur le spectre sont : F_i et E_i pour i = 1,2,3, où F_i est la fréquence en Hertz, E_i l'énergie en dB du $i^{ième}$ formant.

RESULTATS _

1. Le plan (F_1 , F_2) et les cas de dégénérescence.

1.1. Le problème de l'énergie formantique.

Au plan perceptif, les expériences telles que celles de (BECKMANS,R.,1977) semblent indiquer que les énergies E_1 et E_2 n'ont pas de pouvoir discriminant important.

Ceci peut sembler contradictoire avec l'importance accordée par les phonéticiens au trait grave/aigu puisqu'il repose apparemment sur le dosage de l'énergie dans le grave et dans l'aigu. Mais les indices participant à ce trait sont difficiles à cerner et varieraientselon la classe du phonème envisagé. (voir par exemple ROSSI,M, 1977).

Les diverses observations que nous avons nous-mêmes effectuées en vue de la correction de F_1 et de F_2 par leurs énergies relatives n'ont pas été concluantes pour les voyelles : il n'apparait pas significativement de possibilités discriminantes nouvelles par rapport à celles apportées par F_1 et F_2. Bien entendu, l'énergie apportée par F_3 et F_4 est à considérer séparément. Quant au cas des consonnes,il se présente différemment.

1.2.Les voyelles dans le plan (F_1 , F_2).

Ayant constaté des différences notables dans les résultats communiqués par les divers chercheurs (DELATTRE, P, 1966, STRAKA, G., 1965, MRAYATI,M, et CARRE, R., 1976, par exemple) sur ce point, s'agissant de voyelles du français, nous croyons utile de donner leur répartition (et non pas uniquement un représentant idéal) pour le locuteur étudié.

Les coordonnées formantiques ont été relevées en se conformant aux deux principes suivants :

1) F_1, F_2 sont relevés dans la partie stable de la voyelle

2) en présence d'une consonne ou semi-voyelle à influence forte (/j/,/k/, /ʁ/ surtout) une voyelle faible,(c'est-à-dire courte) n'est pas retenue. (Nous reviendrons sur ce problème plus loin).

La figure 1 rassemble les observations ainsi retenues. Des frontières linéaires séparent les diverses régions.Sans la contrainte 2) environ 10 % d'erreurs apparaîtraient : /ɛ↔ e/,/e↔ i/,/œ ↔ɛ/,/ɔ↔ œ /,/ɑ↔ ɛ ,ɔ /,/i↔ y/,/o ↔u/.

Nous avons fait figurer les phonèmes ayant un seul formant sur la droite F_1=F_2. En réalité il s'agit,la plupart du temps,de phonèmes à deux formants voisins réunis en une seule masse dans le spectre. Cela explique la localisation comprise entre F_1 et F_2 du phonème à deux formants correspondant.

Parfois le formant F_2 disparait, probablement absorbé par F_3 et F_4. Dans ce cas,il faut prévoir un traitement approprié, déclanché lorsque le couple de coordonnées ne tombe pas dans la région normale pour (F_1 , F_2). Les expériences perceptives (LANDERCY, A. et RENARD, R., 1975) semblent indiquer que dans ce cas le phonème est perçu comme un /i/. Les observations faites ne nous donnent que des cas de ce genre. Notons aussi certains cas d'ambiguité causés par l'absence de F_2 dans le /u/ et dans le /o/. Le troisième formant, s'il est interprété comme F_2, provoque la confusion avec /œ/,/ø/,/ɛ/ou /e/. Pour éliminer cette source d'erreur il faut classer le phonème comme compact si F_2 est trop faible relativement à F_1.

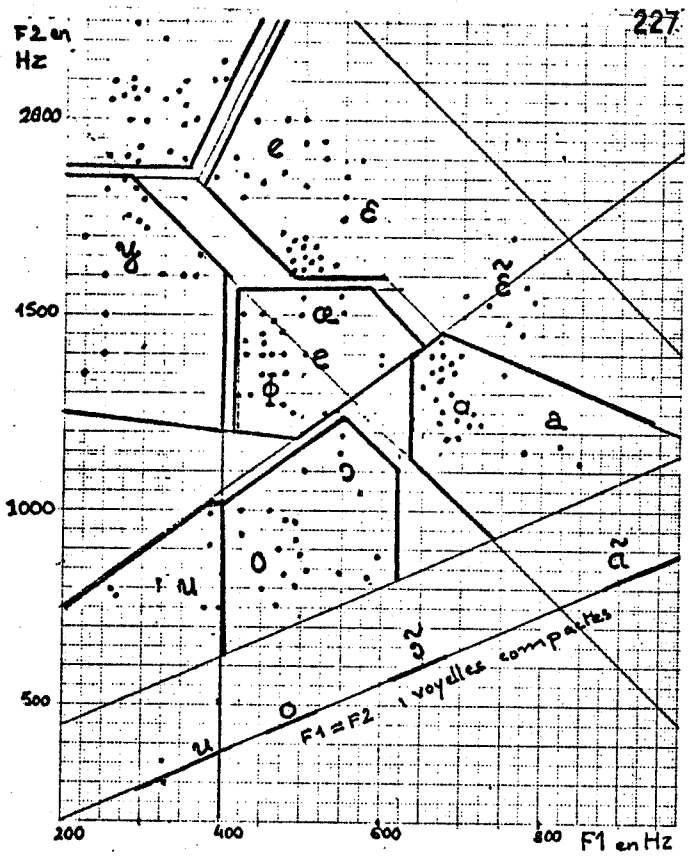


Fig. 1
Voyelles françaises dans le plan (F_1,F_2) prelevées dans 19 phrases prononcées par un locuteur.

French vowels in (F_1 ,F_2) plane taken out of 19 sentences uttered by one locutor.

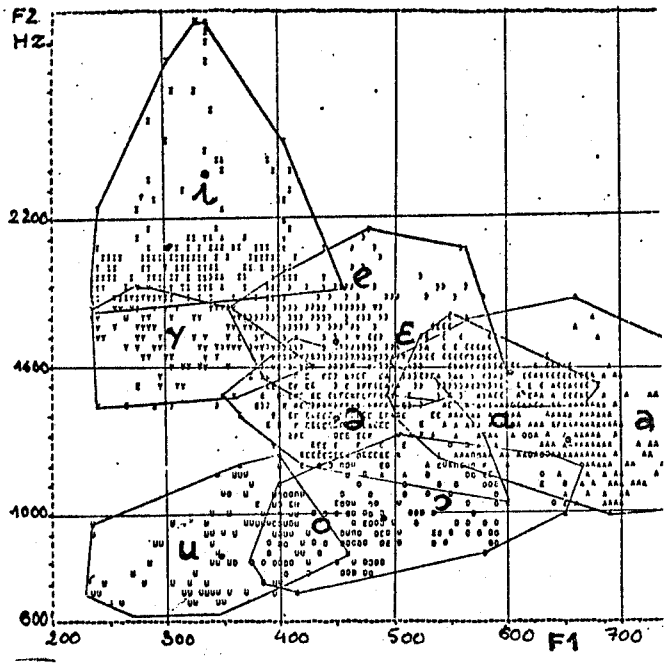


Fig. 2
Voyelles d'un locuteur dans le plan (F_1, F_2) lorsque les fréquences sont prélevées à toutes étapes du phonème.

Vowels of one locutor in the (F_1 , F_2) plane if the frequencies are taken out at every all times.

1.3. Utilisation de F_3 et F_4

Notons enfin que les résultats seraient notablement dégradés si les relevés n'étaient pas effectués dans la région de stabilité de la voyelle (voir fig. 2).

Lorsque le deuxième et le troisième formant coexistent, ce dernier devient significatif pour discriminer dans le complexe /i/,/y/,/e/. Plusieurs procédés peuvent s'envisager :

a) décider /i/ si $F_3 > 2600$, sinon décider /e/ si $F_3 > 2250$, sinon décider /y/ (voir à titre d'exemple la figure 2) ;

b) faire une moyenne pondérée de F_2 et F_3 en posant $F_3 = F_2$ si le troisième formant n'est pas apparent. On obtient ainsi un deuxième formant F_2' corrigé ;

c) dans (PERENNOU, G., DOURS, D., FACCA, R., 1975) on propose la correction suivante de F_2

$$F_2' = \Delta_2 + F_2$$

$$\Delta_2 = 2(F_3 - B)_+ (0{,}45 - F_1)_+ \quad (\text{ou } (x)_+ = \begin{cases} x \text{ si } x \geqslant 0 \\ 0 \text{ si } x \leqslant 0 \end{cases})$$

$$B = 2{,}4 - 0{,}4(F_1 - 0{,}2), \text{ si } F_1 < 0{,}45$$
$$B = 2{,}3 \qquad\qquad \text{si } F_1 > 0{,}45$$

Cette correction fait passer la région de confusion /i ↔y/ de 15 % environ à 1% au plus pour le locuteur et l'échantillon examinés.

Ceci est encore insuffisant en pratique pour rendre compte de nombreux cas où le 3ème formant est pratiquement inexistant relativement à F_4. Si la fréquence haute est supérieure à 3200 Hz (seuil variable selon le locuteur), elle concer-

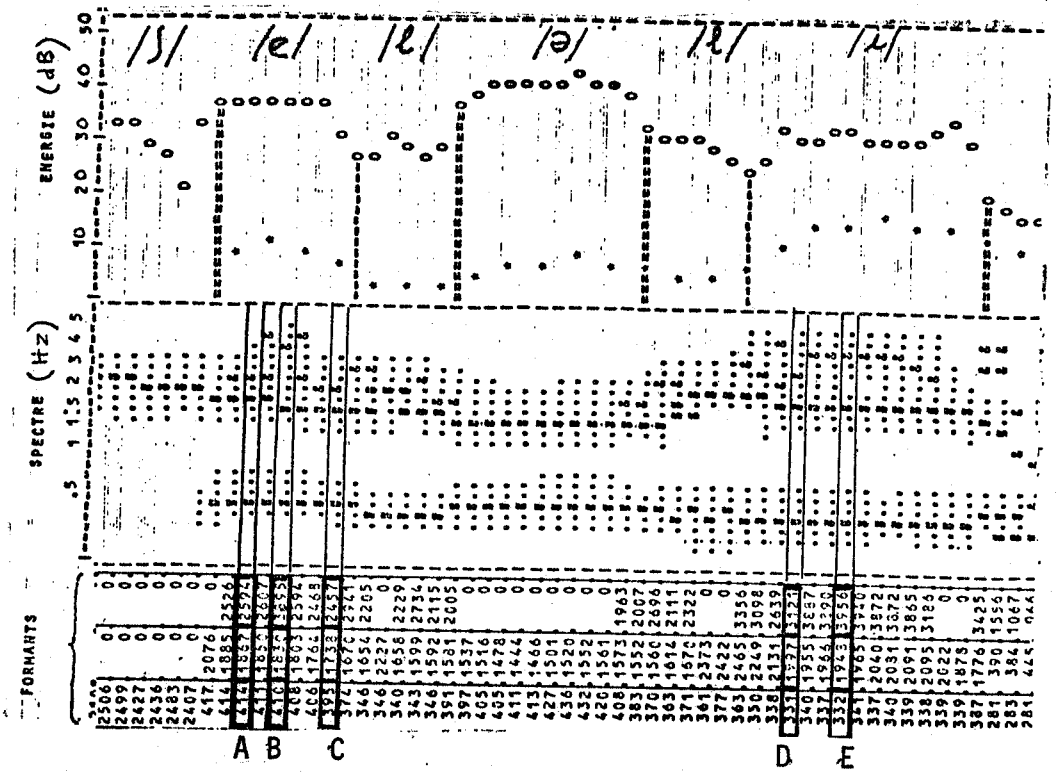


Fig. 3

Pour le phonème /e/ en t_1 et t_2 il y a confusion possible avec /i/. Pour le phonème /e/ on a F_3=2594 en A, F_3=2457 en B, F_4=3695 en C, ce phonème sera reconnu comme /e/ sans ambiguité. Pour le phonème /i/ on a F_3=3121 en D et F_4=3956 en E. Ce phonème ne peut donc pas être confondu avec /y/.

A confusability with /i/ may occur for the phoneme /e/ in t_1 and t_2. For the phoneme /e/ we obtain F_3=2594 in A, F_3=2457 in B, F_4=3695 in C. This phoneme shall be identified as /e/ without ambiguity. For the phoneme /i/ we obtain F_3=3121 in D and F_4=3956 in E. Then, it cannot be possible to have a confusability between this phoneme and /y/.

---

ne F_4. Dans ce cas : si $F_4 > 3750$ Hz décider /i/ sinon décider /e/ (le seuil 3750 Hz est en fait à adapter au locuteur). (Voir figure 3).

2. Influence du contexte sur la partie stable d'une voyelle.

Pour la reconnaissance automatique de la parole la tendance est de faire la reconnaissance d'une voyelle sur sa partie stable. Cela sous-entend que les autres parties sont influencées par les phonèmes voisins (effet de coarticulation). Il faut pourtant reconnaître que cela ne nous met pas à l'abri d'erreurs dans la reconnaissance de ces phonèmes.

Observons à cet égard que les expériences de perception (voir notamment LANDERCY A., et RENARD R., 1975) montrent que si l'on prélève les voyelles des mots où elles s'insèrent, leur reconnaissance est loin d'être excellente.

Par contre, les mots tout entiers sont bien mieux reconnus. Signalons à cet égard les résultats de (KLATT, D.M., et STEVEN K.N., 1973). On peut en tirer diverses conclusions. Une première est que les contraintes lexicales diminuent l'incertitude puisque toute combinaison de sons n'est pas possible a priori.

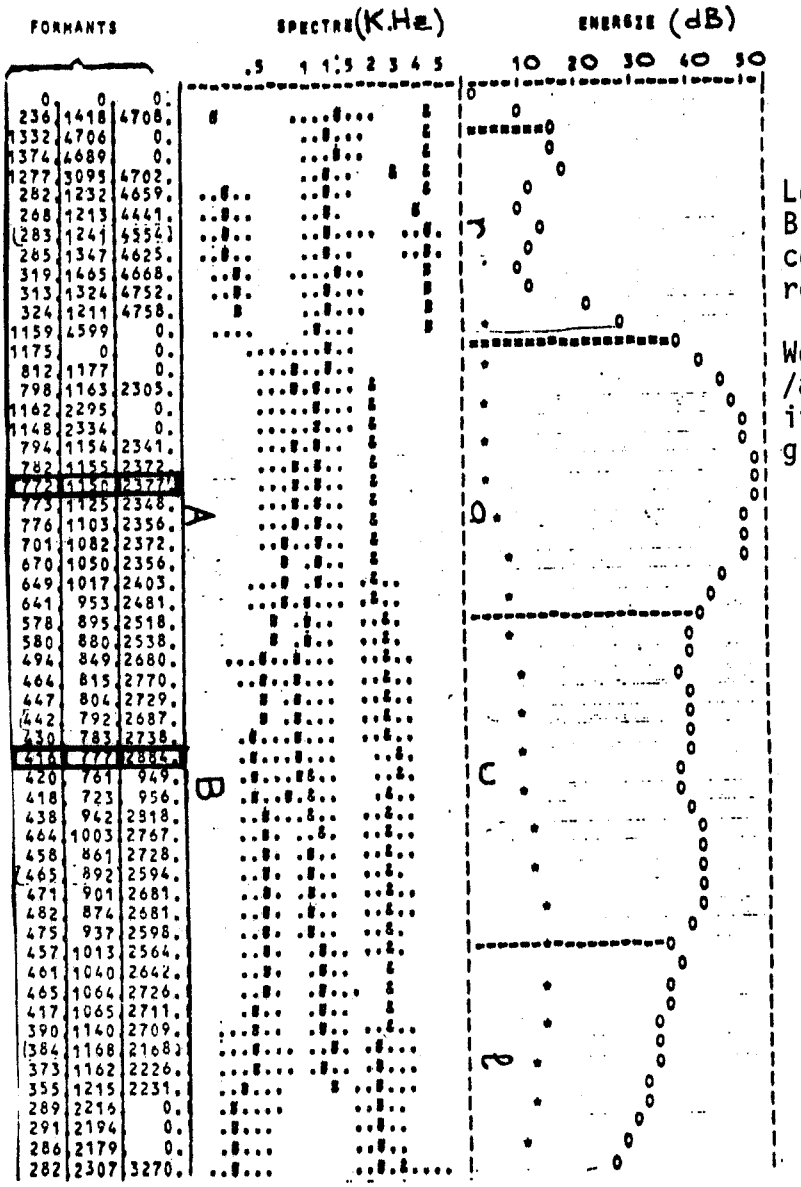


Fig. 4

Le contraste /a/ en A et /u/ en B est suffisant et au-delà de ce qui est nécessaire pour réaliser /a/ /ɔ/ ou /a//o/.

We observe sufficient contrast /a/ in A and /u/ in B ; moreover it is more than required to give /a//ɔ/ or /a//o/.

Il reste que même pour des mots artificiels (ne figurant pas dans le lexique), éventuellement réduits à une voyelle, nous aurions une meilleure reconnaissance qu'avec les voyelles extraites (si les mots artificiels ne sont pas trop complexes). Nous pensons donc qu'en dehors des contraintes phonologiques et lexicales existent des contraintes phonétiques permettant de diminuer l'incertitude sur les voyelles.Deux types de faits seront illustrés à cet égard.

2.1. Contraste de deux voyelles adjacentes.

Dans le nom propre Raoul, deux voyelles /a/ et /u/ sont adjacentes. Le passage de l'une à l'autre se fait à travers la série /a/,/ɔ/,/o/,/u/.

Si l'on observe le mot tel qu'il a été prononcé, on voit que le /u/ attendu se réalise dans une région d'incertitude /u↔o/.(Voir figure 4).

Doit-on mettre,en reconnaissance automatique,ces deux phonèmes comme phonèmes candidats ?.

La logique voudrait que non. En effet, pour exprimer le contraste,(A,O) (A : archiphonème /a,ɑ/ et O : archiphonème /ɔ,o/),il suffisait de réaliser /a/ puis /ɔ/ et même /o/ relativement ouvert. Ni la facilité articulatoire, ni

l'intelligibilité n'exigent d'aller à la frontière entre /u/ et /o/.

L'autre alternative,que nous retenons, est donc de considérer que le contraste des deux phonèmes est suffisant pour forcer la décision en faveur de /u/.

Le problème se pose dans les mêmes termes pour : /ua/,/jɛ/,/ja/,/ae/.

2.2. Influence spécifique des consonnes adjacentes.

Il convient d'abord d'indiquer que cette influence est potentielle. Elle s'exerce d'autant plus que la consonne est articulée avec énergie et que la voyelle est faible.

Il est possible de classer les influences des consonnes selon le lieu et le mode d'articulation.

a) Les bilabiales se caractérisent par une attraction des formants vers la région située entre 700 à 1000 Hz, les dentales vers la région comprise entre 1500 à 1800 Hz ; enfin les palatales se comportent comme les bilabiales pour /ɔ/,/o/ et /u/ mais ont une zone d'attraction située entre 2000 et 3000 Hz pour les autres voyelles (DELATTRE,P., 1966 et MALMBERG, G., 1971).

b) Les consonnes voisées, plus particulièrement les occlusives et les nasales ont une zone d'attraction située au-dessous de 250 Hz.

- Le /r/ présente une zone d'attraction vers la zone 1200-1500 Hz qui provoque souvent une version compacte de la voyelle.

- Le /l/ a peu d'influence. Cependant, on observe qu'il tend à se démarquer au moyen d'un indice lié au 1er formant (parmi d'autres) : à savoir que celui-ci tend à être plus bas que celui des voyelles adjacentes. Cela s'accompagne souvent par le relèvement de ce même 1er formant des voyelles fermées ou semi-fermées-adjacentes.

- Les/j/ et /ɲ/ ont une influence forte de palatales voisées.

- Nous n'avons pas constaté une grande influence des nasales /m/ et /n/ sur F_2 .

Les effets conjugués des deux types de pôles d'attraction (lieu et mode d'articulation) ont des effets qui,à première vue, peuvent sembler contradictoires.

En figure 5 est illustréel'influence des nasales /m/ et /n/ pour le locuteur étudié.

Peut-être peut-on avancer que si F_2 est au-dessus de 500 Hz il est plus influencé par le pôle d'attraction lié au lieu d'articulation ?

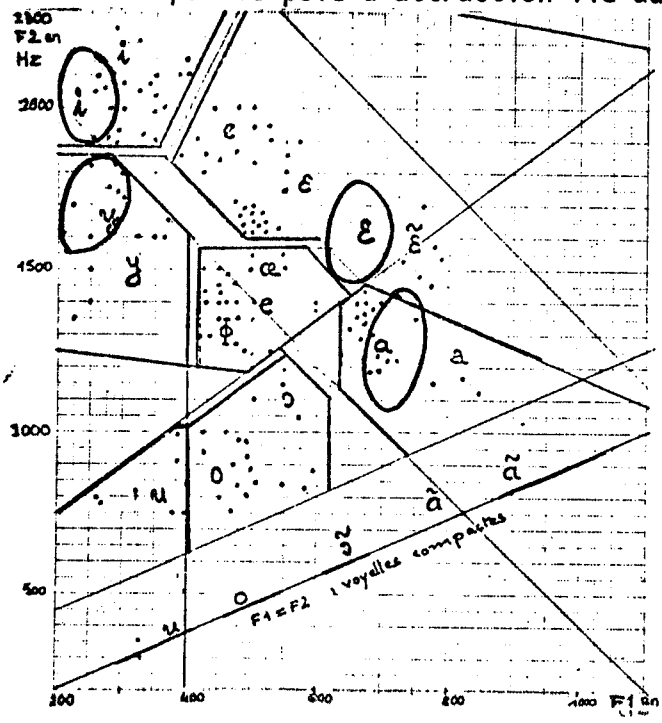


Fig. 5
Localisation de /i/,/y/,/ɛ/,/a/ en contact avec /m/ et /n/.
F_1 est abaissé pour /i/ et /y/ dans le contexte (+Nas).Pour/ɛ/et/a/F_1 est au contraire augmenté.

Localization of/i/, /y/, /ɛ/,/a/ in contact with /m/ and /n/.
In the context (+ Nas) , F_1 is lowered for /i/ and /y/. On the other hand, for/ɛ/and /a/F_1 is increased.

3. Influence des consonnes adjacentes sur la dynamique des voyelles.

Nous nous bornerons à mentionner ce point, pourtant essentiel à la reconnaissance de la syllabe.

On devrait traiter ici d'indices ayant trait à l'évolution des formants en fonction des consonnes adjacentes afin de caractériser leurs lieu et mode d'articulation. On devrait également y caractériser les semi-voyelles.

CONSEQUENCES EN RECONNAISSANCE DE LA PAROLE

Le plan (F_1, F_2), les indices liés à F_3, F_4 et à la compacité permettent d'étiqueter les blocs consécutifs de 8 ms (dans notre système).

Lorsqu'il y a plusieurs phonèmes candidats sur un segment il est possible de lever quelques ambiguités au niveau syllabique. Donnons-en le principe dans les grandes lignes.

Formons un tableau C dont le terme c(i,j) exprime l'influence de la consonne i sur la voyelle j au moyen d'un couple (d_1, d_2) : $d_1 \in \{+,o,-\}$ et $d_2 \in \{+,o,-\}$ d_i = + signifie l'augmentation de F_i, d_i =-la diminution de F_i, F_i = 0 pas d'influence déterminée.

La construction d'un tel tableau pour un locuteur donné exige beaucoup de signal parole analysé. Nous pouvons néanmoins indiquer des résultats assez bien confirmés :

C(m,a) = (+,-) , C(j,a) = (+,-)
C(n,a) = (+,o) , C(ɲ,a) = (+,-)
C(m,E) = (+,-) , C(m,i) = C(n,i) = (-,-)
C(n,ɛ) = (+,o) ,
C(OCCL.VOIS,VOY.NON OUV) = (-,)
C(r,{a,e,ɛ}) = (+,-)
C(r,{ ,o}) = (+,o)

Soit à examiner, par exemple, la suite des phonèmes candidats suivants :

ɔ(+,)	n	ɛ(+,o)
a(-,)	m	a(-,+)
(frontière ɔ,a)		œ(+,)

Nous pouvons alors par consultation de C écarter /a n/ et /a m/ puisque C({m,n},a) = (+,).

La suite /ɔ n ɛ/ est ensuite retenue devant /ɔ n œ/

La procédure permet ainsi de classer deux suites de trois phonèmes et d'en écarter 10 (dans un premier choix).

CONCLUSION

L'analyse du signal par un modèle d'oreille nous permet d'exploiter au mieux les indices formantiques pour l'étiquetage des blocs.

Un certain nombre d'ambiguités restantes semblent devoir être levées au niveau de la syllabe dans le plan (F_1, F_2) sous l'influence spécifique des autres phonèmes. Celle-ci, bien que complexe, semble conforme à une logique phonétique. Si cela se confirmait, il serait possible de réaliser plus facilement l'adaptation au locuteur.

REFERENCES

BEECKMANS, R., 1977, Une approche du processus d'identification de voyelles en français : influences comparées de plusieurs variables expérimentales sur l'espace perceptif. R.A. Institut de Phonétique, Bruxelle

CAELEN, J., 1974 : un modèle mathématique de cochlée et son application à l'analyse de la parole. Thèse de Docteur-Ingénieur, Toulouse.

CAELEN, J., EL JAI, M.C., PERENNOU, G., 1977 : Méthodes structurelles en reconnaissance de la parole au niveau acoustique inspirées par un modèle d'oreille. C.R.E.T.S. - Nice.

CALBRIS, 1974 : Test sur la perception des voyelles orales françaises par des français et des étrangers. Phonetica - vol. 30 n° 2.

DAVID, BROAD and HISASHI WAKITA, 1977, Piecewise-planar representation of vowel formant frequencies. J.A.S.A. , vol. 62 n° 6.

DELATTRE, P., 1966 : Studies in french and comparative phonetics. Mouton & Co London. The Hague. Paris.

KLATT, D.M. et STEVENS, K.N., 1973 : On the automatic recognition of the continuous speech : implications from the spectrogram reading experiment. I.E.E.E. Trans. Acoust. Audio., vol. AU.21. June

LANDERCY, A., et RENARD, R., 1975 : Perception des voyelles françaises filtrées. 6èmes Journées d'Etude sur la Parole- Toulouse.

LIENARD, J.S., 1977 : Les processus de la communication parlée. Masson.

LINSQUIST, GAUFFIN et SUNDBERG, 1976 : Acoustic properties of the nasal tract. Phonetica. vol. 33 n° 3

MALMBERG, B., 1971 : Les domaines de la phonétique. Collection Sup. Presses Universitaires de France.

MRAYATI, M., et CARRE, R., 1976 : Relations entre la forme du conduit vocal et les caractéristiques acoustiques des voyelles françaises. Etude des distributions spatiales. Vol. 33 n° 4.

PERENNOU, G., DOURS, D., FACCA, R., 1975 : Analyse temporelle du signal vocal, comparée à l'analyse fréquentielle du point de vue de la reconnaissance. Projet SESORI IV/137 n° 74-94.

ROSSI, M., 1977 : Les traits acoustiques. La linguistique, vol. 13, fasc. 1.

SANTERRE, LAURENT, 1972 : Transitions articulatoires dans la parole réelle. 3èmes Journées d'Etude sur la Parole. Lannion.

SERNICLAES, R., 1975 : Traitement indépendant ou interaction dans le processus de structuration perceptive des indices de voisement. R.A. Inst. de Phonétique, Bruxelle. 9/2.

SERNICLAES, R., 1975 : Prévoisement et délai d'établissement du voisement : occlusives. R.A., 10/1. pp. 83-104.

STRAKA, G., 1965 : Album phonétique. Les Presses de l'Université de Laval. Québec.

WAJSKOP, M., 1971 : Identification des occlusives inter-vocaliques en français. pp. 102-111. R.A. n° 5. Septembre.

WAJSKOP, M., and SWEETS, J., 1973 : Voicing cues in oral stop consonnant. Journal of Phonetics 1-121-130.

9èmes JOURNEES D'ETUDE SUR LA PAROLE

LANNION 31 mai - 2 juin 1978

ANALYSE DE VOYELLES AVEC DES METHODES DIGITALES

T. de Graaf

Institut des Sciences Phonétiques
Université de Groningue
Pays-Bas.

.**RESUME**

La contribution ci-présente décrit le programme d'analyse de la parole pour la détermination des indices acoustiques de voyelles et de diphthongues qui se réalise à Groningue. Dans les méthodes utilisées on applique des transformations de Fourier pour déterminer le spectre de certains segments de parole. Avec une "routine de FFT"ce calcul se fait d'une façon très efficace à l'aide de l'ordinateur, qui prépare aussi les figures et les tables de paramètres. La méthode de la Prédiction Linéaire (LPC) est appliquée pour déterminer les fréquences des quatre premiers formants dans des segments de parole. Les segments choisis sont plus longs que la période de la voix, tandis que le nombre de segments correspond à la durée de la voyelle ou de la diphthongue.

Cette analyse est appliquée à la description acoustique des indices de son, en particulier des deux formants qui caractérisent les voyelles et les diphthongues de la langue frisonne. Les diphthongues peuvent être caractérisées par deux paires de formants, une pour le segment initial et l'autre pour le segment final de la diphthongue. Les résultats seront représentés dans une Table et nous en donnerons des illustrations dans quelques Figures.

SUMMARY

In this contribution we describe the digital signal analysis techniques which are applied in the Groningen Institute of Phonetic Sciences in order to study the acoustical parameters of speech, in particular for vowels in the Dutch and the Frisian language. The configuration that is used for this analysis is represented in Figure 1. We use the memory of the B&K 7502 digital event recorder in order to store speech segments in 10240 words of 8 bits on disk files, from where they can be elaborated in various ways. The input sampling frequency is taken as 20 kHz in order to reproduce the spectral components between 0 en 5000 Hz with the appropriate intensity. Through the digital event recorder we can recover the analog signal and reproduce its oscillogram on the B&K 7502 level recorder.

In the oscillogram, particular segments are chosen which are further analysed, e.g. by determining the spectrum. For this purpose we use a FFT procedure in the minicomputer and produce spectral plots on the VERSATEC electrostatic lineprinter. In Figure 2 we reproduce an example of spectra which are obtained in this way for two parts of the diphthong [ɔu], one 60 msec and one 250 msec from the initial point of this speech sound. If we take a number of consecutive segments, we get an impression how the spectral composition changes during the pronunciation of this and other diphthongs.

A second way to characterize the change in quality of speech sounds is provided by a Linear Prediction analysis. Here the prediction coefficients are calculated in the autocorrelation method and from these coefficients we determine the first four formant frequencies in a number of speech segments, which follow each other at distances of 12.8 msec. Figures 3 and 4 give examples of the results which are obtained for a number of Frisian vowels and diphthongs.

The Frisian language has a rich system of vowels and diphthongs, which have not yet been properly described in terms of acoustic parameters. The articulatory parameters have been given in a recent book on Dutch and Frisian phonology, where 30 vowels and diphthongs are listed. We used this list in order to investigate the Frisian vowel system. Isolated words containing these vowels and diphthongs in the context w V t were spoken by a native speaker and registered on disk files. From these digital signals the data on the formants were obtained by the LPC analysis. We find that most diphthongs have a characteristic pattern which consists of three parts: an initial vowel-like part with constant formant frequencies, a short transition part ($\lesssim$100 msec) and a final part with again a more or less constant formant pattern. In Table I the diphthongs are characterized by the averages of the first and second formant frequency in the initial (F1, F2) and in the final part (F1′, F2′). Most vowels have a constant formant pattern, whereas a number of long vowels show slow changes in the values of F1 and F2.

ANALYSE DE VOYELLES AVEC DES METHODES DIGITALES

T. de Graaf

Institut des Sciences Phonétiques
Université de Groningue
Pays-Bas.

INTRODUCTION

Pour obtenir l'identification de voyelles d'une langue et donner leur description il est possible de donner des paramètres acoustiques et de les représenter dans des diagrammes. A l'Institut des Sciences Phonétiques de l'Université d'Amsterdam on a directement déterminé des formants avec des oscillogrammes qui représentent la parole. De cette façon on a étudié certaines propriétés des voyelles et des diphthongues dans la langue néerlandaise (MOL, H., 1969; KOOPMANS - VAN BEINUM, F.J., 1969). A l'Institut de l'Organisation RVO-TNO de Soesterberg, les voyelles néerlandaises ont été décrites d'une façon plus générale par des points dans un espace multi-dimensionel; en utilisant des filtres de 1/3 octave on a déterminé des intensités à plusieurs fréquences (POLS, L.C.W., 1977).

En continuant cette tradition néerlandaise, l'Institut des Sciences Phonétiques de Groningue s'occupe de la détermination de propriétés acoustiques de voyelles et de diphthongues en appliquant des méthodes digitales. Dans la contribution ci-présente nous décrivons le système qui a été développé à cet Institut pour faire la recherche phonétique à l'aide d'un ordinateur.

Nous présenterons quelques applications qui font partie d'une analyse générale des traits phonétiques de la langue néerlandaise et du frison, la langue que l'on parle dans une des provinces du nord des Pays-Bas et qui a un système de voyelles et de diphthongues très élaboré.

PROCEDURE EXPERIMENTALE

Dans la Figure 1 nous représentons la configuration qui est utilisée pour l'analyse digitale des signaux. Les segments de parole sont introduits dans une mémoire digitale du type B&K 7502 ("digital event recorder"), qui est utilisée pour la reproduction de la parole comme signal auditif ou comme oscillogramme. Après le contrôle visuel et auditif le segment de parole est transporté à l'ordinateur et enrégistré sur le disque comme un signal digital de 10240 mots de 8 bits. En général nous prenons des échantillons à des distances de T= 0.05 msec (F_S= 20 kHz); la plupart des voyelles ne sont pas plus longues qu'une demi-seconde.

Le signal analogue doit être filtré pour éviter les effets d'"aliasing". Cela veut dire que les fréquences au-dessous de 5000 Hz sont bien représentées, tandis que les fréquences plus hautes sont atténuées. La représentation du signal est optimale, si l'intensité est réglée de sorte que le voltage maximum monte jusqu'à 5.00 Volt.

Sur l'oscillogramme qui est produit par le "level recorder" du type B&K 2305 nous pouvons indiquer les segments du signal qui doivent être analysés. On peut montrer ces segments sur l'oscilloscope et fixer leur points initiaux et leur longueur avec une précision de 0.05 msec. Ensuite les programmes d'analyse digitale peuvent élaborer ces segments de plusieurs façons: On peut créer d'autres signaux par une certaine déformation, les introduire dans le B&K 7502

et faire des stimulus en vue d'expériences psycho-acoustiques. D'autre part il est possible de faire des analyses spectrales et de déterminer des indices acoustiques du signal. Le résultat peut être représenté sur l'écran ou imprimé sur l'appareil du type VERSATEC, avec lequel on peut produire des figures ou des tables qui seront présentées.

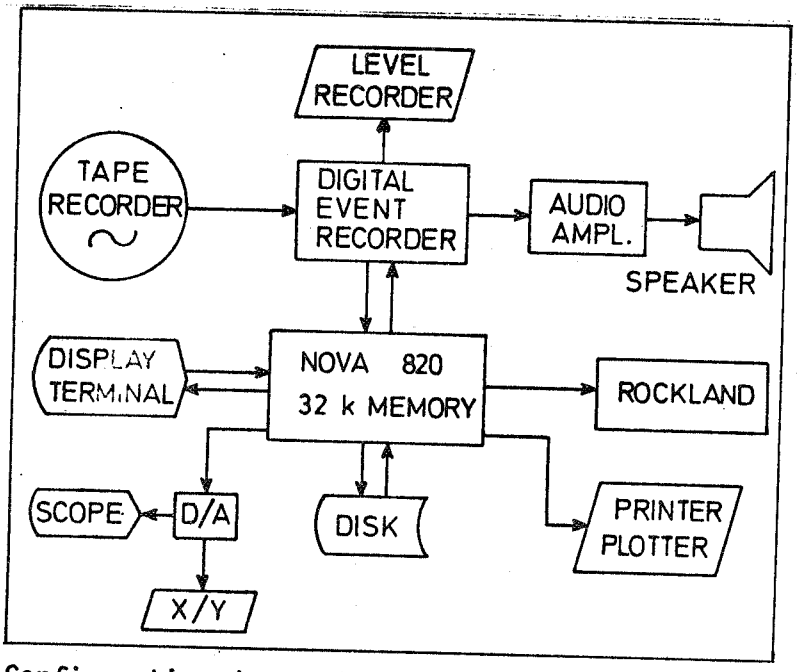


Fig. 1 Configuration du système digital pour l'analyse de la parole à l'Institut des Sciences Phonétiques de Groningue.
Configuration of the Digital Speech Analysis System at the Institute of Phonetic Sciences, Groningen.

METHODES DIGITALES POUR L'ANALYSE DE LA PAROLE

Dans le programme de l'Institut des Sciences Phonétiques de Groningue nous utilisons plusieurs méthodes pour l'analyse de la parole, dont le FFT (Fast Fourier Transform) et le LPC (Lineair Predictive Coding) sont les plus importantes.

Avec la transformation de Fourier un segment de 1024 points est transformé en une fonction complexe de la fréquence, avec laquelle on peut determiner le spectre. En utilisant $T = 0.05$ msec et $F_s = 20$ kHz cela veut dire qu'on analyse un segment de 50 msec, une partie du signal prise comme un segment stationnaire. Les détails de cette méthode ont été décrits dans la littérature (BRIGHAM, E.O., 1974; RABINER, L.R., & GOLD, B., 1975; OPPENHEIM, A.V., & SCHAFER, R.W., 1975). On obtient le spectre en combinant les points réels et les points imaginaires de la fonction complexe. Les premiers 256 nombres correspondent au spectre entre 0 et 5000 Hz. Ce spectre est imprimé par le VERSATEC; nous en donnerons un exemple dans la Figure 2.

La Prédiction Linéaire est une technique très élaborée si l'on veut déterminer des indices acoustiques de la parole, comme la fréquence fondamentale et les formants de voyelles (MARKEL, J.D., & GRAY, A.M., 1976). Nous déterminons les quatre premiers formants dans des segments qui sont plus longues que la période de la voix ("pitch asynchronous analysis"). On peut alors déterminer les coefficients de la prédiction en utilisant la méthode d'autocorrelation. L'ordre de la prédiction M est pris comme 20 pour obtenir un polynôme en z^{-1}, dont les racines déterminent la position des formants et leur largeur.

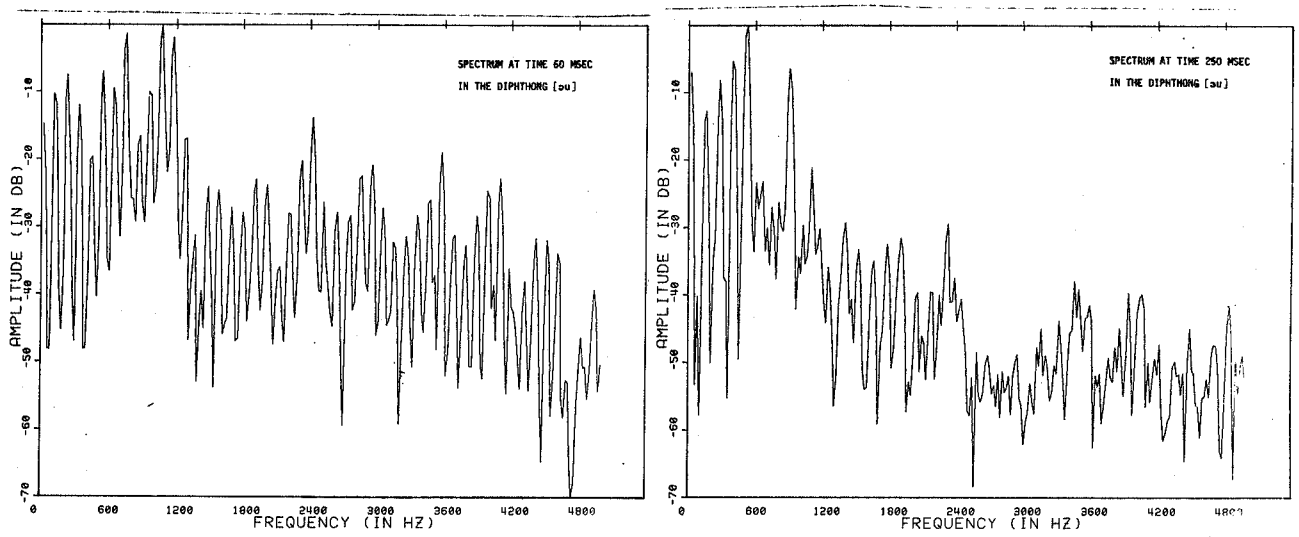


Fig. 2 Spectre de la première et dernière partie de la diphthongue [ɔu]
Spectrum for the first and the last part of the diphthong [ɔu]

Sur l'oscillogramme nous déterminons les points qui correspondent au commencement et à la fin du segment que nous voulons analyser (par exemple une diphthongue) et ensuite nous déterminons les quatre premiers formants F1, F2, F3 et F4 dans les segments consécutifs entre ces deux points. Nous prenons une distance de 256 points entre les points initiaux, qui correspond à 12.8 msec par segment. La figure produite par le VERSATEC donne une idée de la façon dont les formants se transforment pendant la prononciation de cette diphthongue. Le résultat d'une telle analyse sera représenté dans les Figures 3 et 4.

DESCRIPTION PHONETIQUE D'UNE LANGUE

Avec les méthodes digitales décrites ci-dessus nous avons analysé le système des sons dans la langue néerlandaise et dans la langue frisonne, qui est parlée par à peu près 400.000 habitants de la Frise, une province dans le Nord des Pays-Bas. Au début de ce siècle, Eykman a présenté une description phonétique des sons de la langue frisonne (EYKMAN, L.P.H., 1907). Dans son livre il donne une classification de sons qui est basée sur une description articulatoire. Le dialecte frison qu'il a étudié possède dans son système 9 voyelles brèves, 2 voyelles moyennes, 10 voyelles longues, 2 diphthongues brèves, 13 diphthongues moyennes, 1 diphthongue longue, 3 triphthongues ordinaires et 1 triphthongue longue. Même pour une langue germanique c'est un système très complexe et cela vaut la peine de l'étudier, également du point de vue acoustique.

Dans le livre le plus récent qui soit décrivant la phonologie du Néerlandais et du Frison (COHEN, A & al., 1972) les auteurs présentent une liste de phonèmes des deux langues dans laquelle l'on voit que le Frison a beaucoup plus de diphthongues que le Néerlandais, qui en a trois: [ɛi], [ʌü] et [ɔu]. En outre Cohen et al. distinguent dans le Frison encore les diphthongues [ai], [ɔ̀i], [iə], [üə], [uə], [ɪə], [ʌə] et [ɔ̃ə]. Dans notre travail nous avons étudié ce système de voyelles et de diphthongues; la représentation de ces sons est la même que dans le livre de Cohen et al.

RESULTATS ET DISCUSSION

Les voyelles et les diphthongues sont prises dans des segments de 0.5 sec qui forment le centre dans des mots w V t. De cette façon elles sont toutes prononcées dans le même contexte comme des mots isolés et accentués. Le signal digital de 10240 mots de 8 bits qui correspond à une demi-séconde de parole est placé sur le disque et ensuite l'on fait l'oscillogramme. Dans l'oscillogramme on choisit des segments typiques de 1024 points qui peuvent être analysés. Dans la Figure 2 nous donnons les spectres qui représentent le son à 60 msec et à 250 msec du commencement de la diphthongue [ɔu]. On voit que la composition spectrale se transforme pendant la prononciation de la diphthongue. C'est la conséquence de l'articulation, dont le changement correspond à la variation des paramètres acoustiques de la parole. Si l'on veut observer cette transformation plus en détail, il est possible de prendre un nombre de segments consécutifs et d'en reproduire le spectre.

Dans le spectre on peut distinguer les harmoniques de la fréquence fondamentale avec des intensités qui sont déterminées par les résonances des cavités supralaryngales. Il est difficile de déterminer ces résonances (c'est-à-dire les formants) à l'aide de cette figure; la méthode LPC est plus efficace pour y parvenir. Avec cette méthode nous avons obtenu des figures qui représentent la position des quatre formants F1, F2, F3 et F4 comme une fonction du temps (Fig. 3 et 4).

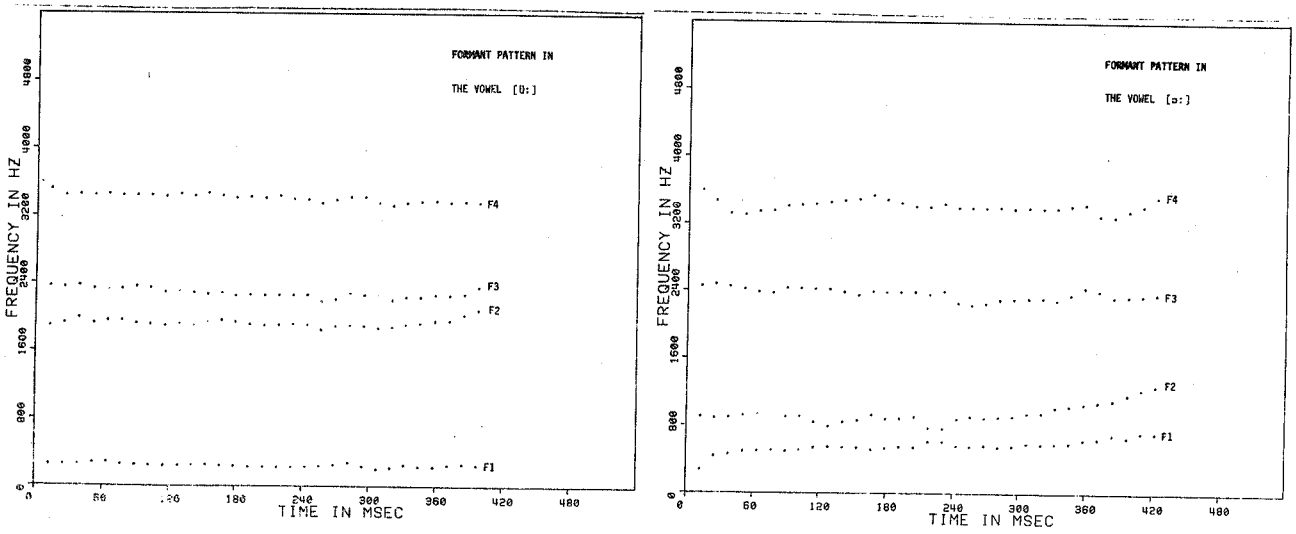


Fig. 3 Les formants F1, F2, F3 et F4 pour quelques voyelles.
The formant frequencies F1, F2, F3 and F4 for a number of vowels.

Nous avons produit des figures semblables pour toutes les voyelles et les diphthongues qui sont énumérées dans le livre de Cohen et al. Dans la plupart des voyelles les formants restent stable pendant toute la durée du son, comme dans le [ü:]. Par contre, les diphthongues montrent très distinctement comment se transforment les positions des formants. Comme Mol l'a déjà remarqué (MOL, 1969), on peut distinguer trois phases: le segment initial qui a le caractère d'une voyelle avec des formants plus ou moins constants, un segment court de transition et un segment final qui a de nouveau des formants aux mêmes fréquences. Dans l'exemple de la diphthongue [üə] on peut dinstinguer ces trois parties de la diphthongue entre 0 et 200 msec, entre 200 et 300 msec et entre 300 et 400 msec.

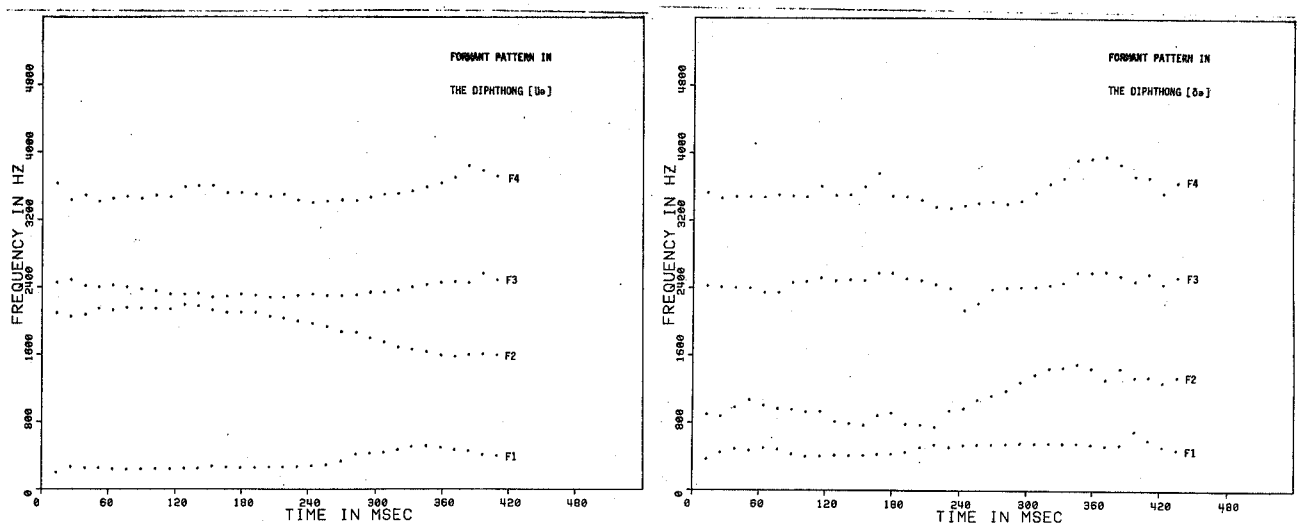


Fig. 4 Les formants F1, F2, F3 et F4 pour quelques diphthongues
The formant frequencies F1, F2, F3 and F4 for some diphthongs

Dans quelques voyelles on peut observer un changement de la position des formants, comme dans le [ɔ:]. Dans ce cas on peut dire que les voyelles ont un caractère diphthongué, bien que l'on ne puisse pas distinguer une transition aussi prononcée que dans [üə]. Dans un autre contexte (par exemple devant le r) ce phénomène est encore plus évident.

Dans la Table I nous donnons les fréquences des deux premiers formants avec lesquelles on peut caractériser les voyelles et les diphthongues. Pour la plupart des voyelles nous avons calculé la moyenne des valeurs que prennent les fréquences dans la domaine stationnaire; pour ce qui concerne les diphthongues et les voyelles diphthonguées, F1 et F2 caractérisent le segment initial, tandis que F1´ et F2´ caractérisent le segment final.

Dans notre recherche nous nous sommes occupés de l'analyse digitale des segments de parole dans un contexte spécial, prononcés par une seule personne comme mots isolés. Il est important d'étudier les traits acoustiques dans des conditions plus générales: en se basant sur les réalisations de plusieurs personnes, en choisissant des contextes plus divers et en utilisant d'autres langues. Le programme pour atteindre ce but forme le cadre d'un projet que nous espérons continuer à l'Institut des Sciences Phonétiques de Groningue.

Table I

Paramètres acoustiques caractérisant les voyelles et les diphthongues frisonnes

	voyelle ou diphthongue	F1	F2	F1′	F2′
1	i	270	2300		
2	i:	240	2350		
3	ü	260	1910		
4	ü:	250	1920		
5	u	260	720		
6	u:	260	760		
7	ɪ	350	2050		
8	e:	360	2140	320	2400
9	ʌ	430	1630		
10	ö:	410	1750	350	1880
11	ð	530	1170		
12	o:	490	980	430	990
13	ə	430	1580		
14	ɛ	550	1840		
15	ɛ:	540	1960	590	1880
16	ɔ	660	970		
17	ɔ:	520	890	590	970
18	a	730	1100		
19	a:	820	1300		
20	ɛi	570	1970	400	2070
21	ʌü	730	1500	440	1740
22	ai	740	1240	480	2160
23	ði	490	940	310	2180
24	ɔu	760	1120	500	970
25	iə	260	2290	470	1620
26	üə	250	2140	460	1650
27	uə	260	750	460	1450
28	ɪə	340	2260	540	1710
29	ʌə	370	1900	540	1590
30	ðə	460	980	550	1430

REFERENCES

BRIGHAM, E.O. , 1974, The Fast Fourier Transform; Prentice Hall, Englewood Cliffs, New Yersey.

COHEN, A. , EBELING, C.L. , FOKKEMA, K. , VAN HOLK, A.G.F. , 1972, Fonologie van het Nederlands en het Fries; Martinus Nijhoff, 's-Gravenhage.

EYKMAN, L.P.H. , 1907, Description phonétique des sons de la langue frisonne parlée à Grouw; Extrait des Archives Teyler, Haarlem.

KOOPMANS-VAN BEINUM, F.J. , 1969, Nog meer fonetische zekerheden; De Nieuwe Taalgids, n° 62, pp. 245.

MARKEL, J.D. , Gray, A.H. 1976, Linear Prediction in Speech; Springer, Berlin.

MOL, H. , 1969, Fonetische zekerheden; De Nieuwe Taalgids, n° 6, pp. 161.

OPPENHEIM, A.V. Schafer, R.W. , 1975, Digital Signal Processing, Prentice Hall, Inc., Englewood Cliffs, New Yersey.

POLS, L.C.W. , 1977, Spectral Analysis and Identification of Dutch Vowels in Monosyllabic Words; Doctoral Dissertation, Amsterdam.

RABINER, L.R. , GOLD, B. 1975, Theory and Application of Digital Signal Processing; Prentice Hall, Englewood Cliffs, New Yersey.

9èmes JOURNEES D'ETUDE SUR LA PAROLE

LANNION 31 mai - 2 juin 1978

ANALYSE D'UNE METHODE PROBABILISTE APPLIQUEE A LA RECONNAISSANCE DES VOYELLES PARLEES

N. Tirandaz et C. Berger Vachon

Laboratoire de Physique Electronique
(Professeur Mesnard)
Université Claude Bernard Lyon

.RESUME

A partir de 7 canaux sélectionnés d'un vocodeur, on cherche à définir une politique probabiliste de reconnaissance des voyelles parlées basées sur l'application de la formule de GAUSS-BRAVAIS.

Dans une première phase dite d'apprentissage, on étudie la répartition des similitudes (probabilités normées sur 7 canaux) des 13 voyelles de la langue française. Pour la reconnaissance on teste 3 politiques possibles :

- reconnaissance à partir de la plus grande similitude
- reconnaissance avec un seuil tenant compte de la répartition statistique des similitudes pour chaque voyelle
- reconnaissance avec seuil très bas.

Ces 3 politiques sont testées avec plusieurs locuteurs et pour la voyelle prononcée isolément ou dans un mot.

Différents aspects sont à envisager :

- le choix de l'échantillon d'apprentissage
- les problèmes posés par l'application de la technique probabiliste
- l'efficacité de la reconnaissance

Les résultats obtenus sont présentés et discutés.

ANALYSIS OF A STATISTICAL CRITERION FOR SPOKEN VOWELS RECOGNITION

N. Tirandaz et C. Berger Vachon

Laboratoire de Physique Electronique
(Professeur MESNARD)
Université Claude Bernard Lyon

.SUMMARY

From 7 selected channels in a vocoder, the anthors wish to define a probabilistic policy for spoken vowels recognition (through the GAUSS BRAVAIS formula utilization).

In a first step called "lerning", the distribution of a similarity (probability averaged upon the 7 channels outputs) is studied for thirteen french vowels.

Three policies are then considered for recognition :

- recognition made upon the greatest similarity
- recognition with a threshold given by the similarity distribution
- recognition with a low threshold

These three policies are tested for several speakers and for vowels spoken separately or in a word.

Several aspects are to be considered :

- the choice of learning samples
- the difficulties raised by probabilistic formula utilization
- the recognition efficiency

Results are presented and discussed

ANALYSE D'UNE METHODE PROBABILISTE APPLIQUEE A LA RECONNAISSANCE DES VOYELLES PARLEES

N. Tirandaz et C. Berger Vachon

Laboratoire de Physique Electronique
(Professeur Mesnard)
Université Claude Bernard Lyon

INTRODUCTION

Nous avons utilisé les quatorze sorties du vocodeur ETA du CNET Lannion qui donnent une paramétrisation fréquentielle du signal vocal. On sait que le signal vocal est très redondant (PIMONOV L., 1965) et lrosqu'on ne dispose que d'un outil informatique limité, cette redondance pourrait être mise à profit pour l'identification des occurences.

BERGER et MESNARD (1976) ont proposé le choix de sept canaux du vocodeur pour la mise en oeuvre d'une reconnaissance probabiliste.

Soit $C = \{C_{\sigma_1}, C_{\sigma_2}, C_{\sigma_3}, C_{\sigma_4}, C_{\sigma_5}, \ldots C_{\sigma_7}\}$ un ensemble de sept canaux, où $\sigma_1, \sigma_2 \ldots \sigma_7$ est une combinaison des C_{14}^7 combinaisons possibles de sept canaux sur quatorze. Le canal C_{σ_p} est corrélé avec les six autres de "l'heptet", avec un taux de corrélation linéaire moyen défini par (BERGER, 1975)

$$\rho_{C\sigma_p} = \frac{1}{6} \sum_{q=1}^{7} \left| \rho(C_{\sigma_p}, C_{\sigma_q}) \right|$$

On définit la qualité d'un canal par :

$$Q_{C\sigma_p} = s_{C\sigma_p} (1 - \rho_{C\sigma_p})$$

où $s_{C\sigma_p}$ est l'aptitude du canal C_{σ_p} pour séparer les voyelles (BERGER et MESNARD, 1976 b)

$$Q_C = \sum_{p=1}^{7} Q_{C\sigma_p}$$

Partant de ces résultats, on choisit les canaux n° 2, 3, 4, 5, 6, 7 et 8 et on essaie d'analyser une méthode probabiliste dans le cas d'une reconnaissance des voyelles.

LA METHODE

L'énergie d'une occurrencepeut être schématiquement représentée par une courbe de la forme ci-dessous : (figure 1)

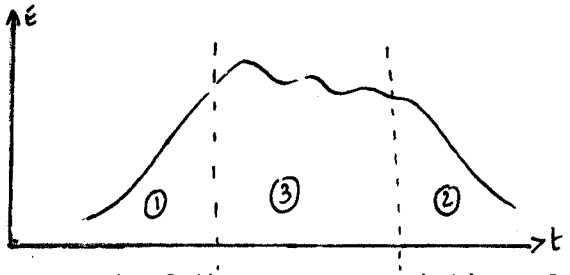


Figure 1 - Schematic aspect of the energy variation of an isolated occurrence
Aspect schématique de la variation de l'énergie d'une occurrence isolée

Schématiquement, on peut distinguer trois zones sur cette courbe :
zone (1) : transition ascendante
zone (2) : transition descendante
zone (3) : zone quasi-stationnaire

On suppose que l'occurrence peut être représentée par un vecteur $\vec{V}$ représentant la zone quasi-stationnaire. Ce vecteur $\vec{V}$ aura pour composante les sept sorties des sept canaux sélectionnés.

En supposant que chaque voyelle est définie par un domaine D_i de l'espace R^7 sur lequel est défini le vecteur $\vec{V}$, il s'agit de décider si le vecteur $\vec{X}$ représentant une occurrence, définit un point M appartenant au domaine D_i d'une voyelle A_i donnée (si tant est que cet espace puisse être parfaitement repéré).

La première étape consistera donc à définir les archétypes. L'archétype A_i est défini statistiquement sur un échantillon de vecteurs le représentant. Sur cet ensemble de vecteurs on calcule : un vecteur moyen $\vec{M}_i$, une matrice de covariance Λ_i le déterminant $|\Lambda_i|$ et Λ_i^{-1}

Dans ces conditions (en supposant que les distributions des énergies sur les canaux ne s'écartent pas trop d'une loi normale), la probabilité pour que le vecteur $\vec{X}$ (représentant l'occurence x) corresponde à l'archétype A_i est donnée par la formule de Gauss-Bravais :

$$\text{Prob}\ (X/A_i) = \frac{1}{(2\pi)^{n/2}\ \Lambda^{1/2}} \exp\left[-\frac{1}{2}\ (\vec{X}-\vec{M})_T\ \ \Lambda^{-1}\ (\vec{X}-\vec{M})\right]$$

Cette formule dépend du nombre de canaux utilisés. Pour plus de commodité, nous avons préféré introduire une similitude $S(X/A_i)$ en normalisant par rapport au nombre des canaux. En effet, la matrice Λ_i n'étant pas toujours inversible , nous avons dû parfois (voyelles e, o, y, ø, i) retirer des canaux pour assurer cette inversion

L'échantillon servant à définir les archétypes est composé de 182 vecteurs répartis en 13 classes (les 13 voeylles de la langue française prononcées 14 fois par le même locuteur). Chaque voyelle est donc représentée par 14 vecteurs.

PHASE D'APPRENTISSAGE ET CALCUL DES SEUILS

On calcule pour chaque occurrence sa similitude avec chacun des 13 archétypes précédemment définis.

La répartition des 182 similitudes pour chaque voyelle va nous permettre de définir des domaines D_i d'appartenance.

Le tableau ci-dessous illustre cette répartition pour la voyelle /a/

Toutes les classes représentant une similitude quasi-nulle ont été supprimées du tableau (chaque classe possédant 14 vecteurs).

X = /a/	.43	.45	.46	.48	.52	.52	.52	.53	.53	.54	.59	.64	.64	.66
X = /æ/										.01	.01	.03	.03	.03
X = /á/											.01	.01	.14	.16

similitude $S(X/A_i)$ quand A_i = /a/
similarity $S(X/A_i)$ when A_i = /a/

La détermination des zones d'appartenance se fait par le choix d'un seuil à partir des 13 tableaux de répartition du modèle ci-dessus. Ce choix obéit en fait à une politique choisie. Trois possibilités s'offrent à nous :

1 - On compare les 13 similitudes calculées et sera reconnu l'archétype pour lequel l'ocurrance x présente la plus grande similitude. Cette politique consiste donc à attacher à toute occurrence un archétype, mais ne donne aucune certitude quant à l'exactitude de l'archétype reconnu.

2 - Choix d'un seuil statistique.

On calcule pour chaque classe, à partir des tableaux de répartitions des S_i, une similitude moyenne S_{im} et un écart-type σ_i et le seuil $s_i = S_{im} - 3\sigma_i$.

3 - On choisit le seuil le plus bas possible, afin que tous les vecteurs appartenant à un archétype soient reconnus, mais on aura alors des déterminations multiples.

Compte tenu des études actuelles sur la redondance du langage, une politique de ce dernier type semble souhaitable.

LA RECONNAISSANCE - RESULTATS

Pour analyser cette méthode probabiliste, nous avons testé les trois politiques précédemment citées dans les cas suivants :

- voyelles isolées ayant servi lors de l'apprentissage
- voyelles prononcées par le même locuteur à l'intérieur d'un mot
- voyelles, isolées ou non, prononcées par d'autres locuteurs

A - Voyelles isolées (182 vecteurs de l'apprentissage + 7 autres)

A.1 - Cas de la plus grande similitude

Le tableau ci-dessous résume les résultats obtenus :

	/a/	/e/	/o/	/y/	/ɛ/	/u/	/œ̃/	/õ/	/ø/	/ə/	/ɑ̃/	/ɛ̃/	/i/
% de reconnaissance	100	100	100	100	100	71	100	100	93	100	100	100	100
confusions						29 /e/		7 /o/					

Reconnaissance voyelle isolée avec la plus grande similitude
Recognition of isolated vowels made upon the greatest similarity

Sur un total de 189 vecteurs, 5 vecteurs sont mal reconnus. Les confusions surviennent entre /ɛ/ et /e/ d'une part et /õ/ et /o/ d'autre part, ce qui est très bon.

A.2 - Cas de $s = S_m - 3\sigma$

Trois possibilités peuvent se présenter :

a) l'occurrence n'est pas reconnue
b) l'occurrence est reconnue pour plusieurs archétypes (RM)
c) le bon archétype n'est pas reconnu

Dans ce cas on ne constate aucune exclusion, ni de non reconnaissance. Il existe cependant des reconnaissances multiples données par le tableau ci-dessous. En observation, nous avons porté le détail de ces reconnaissances multiples (nombre et voyelles)

Recognition with a threshold $s = S_{im} - 3$
Reconnaissance avec le seuil $s = S_{im} - 3$

	/a/	/e/	/o/	/y/	/ɛ/	/u/	/œ̃/	/õ/	/ø/	/ə/	/ɑ̃/	/ɛ̃/	/i/
RM	0	2	7	0	5	3	2	9	0	2	0	0	0
Observations		2 /ɛ/	7 /õ/		5 /e/	3 /œ̃/	1 /u/ 1 /ø/	7 /o/ 2 /ɛ̃/		2 /u/			

(RM : reconnaissance multiple)

A.3 - Cas du seuil très bas

Nous avons testé 72 vecteurs représentant les voyelles /a/ , /e/ , /o/ /u/ , /ɛ/ , /i/ . Dans ce cas, on élimine la possibilité d'avoir des exclusions et des non reconnaissances, mais on augmente les reconnaissances multiples.

Le tableau ci-dessous résume les résultats obtenus pour les confusions.

phonèmes	/a/	/e/	/o/	/ɛ/	/u/	/i/
total RM	3	22	19	4	14	3
détail RM	3 /ɑ̃/	13 /e/ 6 /ø/ 5 /ɛ/	15 /ɔ̃/ 4 /u/	4 /ɛ/	3 /o/ 7 /œ̃/ 1 /ø/ 3 /ə/	1 /e/ 1 /ɛ/ 1 /u/

Reconnaissance avec seuil très bas
Recognition with a low threshold

B - Cas des voyelles dans des mots prononcés par le même locuteur (sur lequel a été fait l'apprentissage)

Nous avons pris les voyelles dans les chiffres.

B.1 - Cas de la plus grande similitude

Tous les /a/ du mot "quatre" sont reconnus comme /u/ .

Le tableau ci-dessous donne les résultats pour les autres. Notre comparaison se limitant aux mots zéro, trois, six, huit et sept, les voyelles /e/ , /o/ , /i/ , /a/ et /ɛ/ .

mots	zéro		sept	trois	six et huit	
phonèmes	/e/	/o/	/ɛ/	/a/	/i/	total
bonne reconnaissance	2	5	7	6	10	30
confusions	1 /ɥ/ 2 /ɛ/ 1 /u/	1 /ɑ̃/ 1 /u/	0	1 /ɑ̃/	1 /e/ 3 /ɥ/	11
	6	6	7	7	14	41

Reconnaissance des voyelles dans les mots avec la plus grande similitude
Recognition of vowels in a word upon the greatest similarity

B.2 - Cas de $s = S_m - 3\sigma$

On n'obtient que onze bonnes reconnaissances sur 41 vecteurs :

mots	zéro		sept	trois	six et huit
phonèmes	/e/	/o/	/ɛ/	/a/	/i/
BR	1	1	0	2	7

Reconnaissance de voyelles dans les mots avec $s=S_{im}-3\sigma$
Recognition in of vowels in a word with $s=S_{im}-3\sigma$

(BR : bonne reconnaissance
good recognition

B.3 - Cas de seuil très bas

Tous les 41 vecteurs sont bien reconnus, on a cependant des reconnaissances multiples.

mots	zéro		sept	trois	six et huit
phonèmes	/e/	/o/	/ɛ/	/a/	/i/
RM	11	0	5	0	0
	6 /ɛ/ 5 /i/		5 /e/		

Résultats avec seuil très bas
Results with a low threshold

On voit donc apparaître de ces triples reconnaissances les 6 vecteurs /e/ de zéro sont reconnus /e/ et /ɛ/ et 5 d'entre eux comme /i/ .

C - Pour d'autres locuteurs

C.1 - Cas de la plus grande similitude

Le tableau ci-dessous donne les résultats obtenus pour 74 vecteurs représentant les voyelles /a/ , /i/ , /ɛ/ , /e/ , /o/ et /u/ .

phonèmes	/a/	/e/	/ɛ/	/o/	/u/	/i/	total
bonne reconnaissance	2	1	1	2	0	10	16
confusions	11 /ɛ/	6 /ɛ/ 2 /y/ 4 /u/	4 /y/ 5 /u/ 2 /e/ 1 /i/	5 /u/ 3 /y/ 1 /õ/ 1 /e/	5 /e/ 1 /ɛ/ 4 /y/	2 /y/ 1 /e/	58
total	13	13	13	12	10	13	74

Results with other speakers

C.2 - $s = S_m - \sigma_3$

Sur 74 vecteurs, 2 seulement sont bien reconnus

C.3 - Seuil très bas

Sur 62 vecteurs, 18 sont bien reconnus.

CONCLUSIONS

Au terme de ce travail un certain nombre de résultats méritent d'être soulignés :

1 - La décision selon la plus grande similitude paraît incompatible avec une politique recherchant à reconnaître les phonèmes par exclusion. Cette méthode est néanmoins très utile en vue d'un classement des phonèmes quand on connaît toutes les possibilités (MERCIER G.,GRESSER J.Y., 1973).

2 - La détermination d'un seuil à partir de la répartition des similitudes calculées sur un échantillon ($1 = S_m - 3\sigma$) donne des résultats intéressants, mais présente l'inconvénient d'une non reconnaissance possible de la bonne classe. Le problème des doubles déterminations ne nous semble pas majeur. En effet, les séquences ultérieures du traitement du discours : probabilité de succession des phonèmes, contraintes morphologiques, sementiques et contextuelles, permettent d'espérer de lever ces incertitudes (HATON J.P, 1974).

3 - En tenant compte de ce qui vient d'être dit, la fixation d'un seuil très bas nous paraît opérationnelle.

4 - L'échantillonnage de base pour la définition des archétypes joue un rôle important et sa constitution mérite une étude particulière.

5 - L'extension de la méthode à d'autres locuteurs demande des procédures d'adaptation.

6 - La similitude que nous avons employée, bien que satisfaisante vis à vis de la formule générale, ne paraît pas rationnelle sur le plan statistique. Il faut en outre remarquer que c'est toujours le même canal (n° 3) qui a empêché l'inversion des matrices en introduisant des covariances nulles. Nous nous proposons donc de ne pas en tenir compte dans le calcul de la similitude. La similitude probabiliste semble un bon moyen pour classer les archétypes mais n'exclut pas l'utilisation conjointe des critères déterministes.

Bien que relativement sophistiqué, le critère que nous avons étudié ne résoud pas le problème du bruit à la sortie de l'étage acoustique. Néanmoins il nous semble intéressant de le retenir, car les méthodes statistiques doivent apporter une aide appréciable pour la classification des sons, lorsqu'elles sont bien utilisées. Elles sont, en outre, complémentaires à beaucoup de méthodes déterministes.

REFERENCES

BERGER-VACHON C. Conception d'une entrée vocale automatique
Thèse d'Etat, 1975, Lyon

BERGER C. et MESNARD G. 1976 a, Application des métriques statistiques à la reconnaissance des zones quasi-stationnaires du signal vocal
7ème journée d'études sur la parole Nancy, 1976

BERGER C. et MESNARD G. 1976 b, Etude de l'erreur introduite par la numérisation du signal dans un problème de reconnaissance des formes : application à la parole. In Annales des Telecommunications
Tome 31, n° 9-10, sept. oct. 1976, p. 321-332

HATON J.P. Contribution à l'analyse, la paramétrisation et la reconnaissance automatique de la parole
1974, Nancy

MERCIER G. et GRESSER J.Y. The automatique segmentation of speech into syllabic and phonetic units, application to french words and utterances. Symposium on auditory analysis and perception of speach. Léningrad, 1973

PIMONOV La parole synthétique et son application dans la correction auditive des sourds
Thèse 1965, Paris

9èmes JOURNEES D'ETUDE SUR LA PAROLE

LANNION 31 mai - 2 juin 1978

LA DUREE DES PHONEMES ET LA RECONNAISSANCE DES DEBUTS ET FINS DE MOTS : ETUDE ACCOUSTIQUE DES STRUCTURES HOMOPHONIQUES CHEZ DEUX LOCUTEURS

GENEVIEVE CAELEN, GEORGES MAURAND - UER DE LINGUISTIQUE

UNIVERSITE DE TOULOUSE-LE-MIRAIL

.RESUME

D'un corpus de 84 phrases énonciatives où sont représentées toutes les structures syntaxiques possibles, on a tiré 67 exemplaires d'expressions phonétiquement semblables mais syntaxiquement différentes. Cette étude, reposant sur une analyse acoustique très précise, au cours de laquelle indices et traits phonétiques ont été pris en compte, propose une description qualitative des " patrons" temporels des unités constitutives de la phrase.

SUMMARY

In order to illustrate a study we explain elsewhere (Thèse de troisième cycle - I978) about the prosodic features (Fundamental frequency, energy and duration) of assertive sentences in french, we consider here the confined sphere of homophonic structures which can offer particulary a good model of syllabic length in words, syntagms and sentences. We purpose a qualitative and quantitative analysis of "time patterns" used by two speakers, a woman and a man.

We define the "tempo" as the succession of syllabic accelerations and slackenings in the sentence, mesured by the average of each unstressed vowel in each consonant environment for each speaker, which enables to make comparisons between these two speakers. They both use the same patterns : inside same syntactic structures, incomplete sense is manifested by progressive slackening during the phonation time of lexical word, syntagm or whole sentence ; the breaking of continuity in the slackenings shows the limits of the semantical unities, regarding to the speaker, which may or not include several lexical words ; every stated unity seems to have the same weight for the speaker, as any other one of similar features. Inside still the same syntactic structures, semantic dependence of one element upon another is marked by a new pattern : the rate of the first one is more and more accelerated, then the second one inverses the movement and comes back to a sensible slackening ; this double tempo points out a semantical hierarchy between the different unities of the assertive sentence.

INTRODUCTION.

A partir d'une analyse non perceptuelle de la durée des phonèmes, consonnes et voyelles, intégrés dans un corpus de 84 phrases énonciatives simples ou étendues*, lues par deux locuteurs des deux sexes, cette étude propose une description acoustique, qualitative et quantitative des "patrons temporels" des unités constitutives de la phrase, pouvant trouver une application à la reconnaissance automatique de la parole.

1 - PROCEDURE EXPERIMENTALE.

84 phrases énonciatives lues et enregistrées en chambre sourde, ont fait l'objet d'une analyse accoustique très minutieuse au cours de laquelle, ont été relevées, en particulier, les réalisations temporelles des voyelles selon les timbres et leur entourage contextuel, ainsi que des consonnes. N'ont été retenues par la suite dans le calcul des moyennes que les voyelles atones ou les consonnes appartenant à des syllabes non proéminentes.

Pour apprécier le "tempo" syllabique, nous avons comparé directement la valeur individuelle de chaque type de phonèmes constitutifs de la syllabe dans les phrases, à la valeur moyenne des phonèmes de même nature (timbre et contexte). Nous avons ensuite additionné en valeur absolue les écarts temporels relevés par rapport à cette moyenne dans le cadre de la syllabe, de chaque voyelle et consonne. Nous obtenons ainsi un tempo phrastique, que l'on peut définir comme la succession des ralentissements (valeurs positives) et des accélérations syllabiques (valeurs négatives) dans la phrase. Cette méthode nous permet de mener une comparaison soutenue entre les réalisations de nos deux locuteurs, dans la mesure où les valeurs moyennes servant de référence sont à chaque fois établies en fonction de leurs propres performances acoustiques ; elle nous permet en outre de nous affranchir très utilement de la complexité de la syllabe (syllabe à 1,2,3,4 ... phonèmes) .

2 - ETUDE DU TEMPO.

2.I Le tempo dans les phrases du corpus :

En étudiant de manière précise le tempo de toutes les phrases du corpus chez les deux locuteurs, on remarque, dans l'utilisation qu'ils en font, une grande régularité. Celui-ci ne varie pas, dans nos réalisations, selon la hiérarchie des différents constituants du texte (mot lexical, syntagme, phrase) ; à tout niveau, on trouve généralement deux choix possibles de réalisation temporelle de ces unités :

2.1.1 Tempo croissant :

Le tempo étant calculé, comme on l'a lu, en prenant une moyenne comme référence (valeur zéro), nous trouvons des "tempos négatifs" à côté de "tempos positifs", calculés en cs :

---

* Voir pour la description syntaxique de ces phrases le volume des actes des 8e J.E.P. (I977) : Geneviève CAELEN et G. MAURAND : Corrélations entre mélodie, intensité et durée dans la phrase française énonciative.

2.1.1.1 Sans rupture de continuité :

. les mots lexicaux :

ex : ɭ ə p ɛ R m i s y o n ɛ R

Locuteur masculin
ou LM : 0 -5 -2 2 25

. les syntagmes :

ex : d ɑ̃ zy n ɛ R n u v ɛ |

LM : -1 -1 5 5 17

. les phrases :

ex : il sõŋ tre zami

LM : -5 -1 2 2 4

2.1.1.2 Avec rupture de continuité :

. les syntagmes :

ex : ɑ̃ ne kutãŋ lə pɛ̃ŋsõŋ

LM : -5 0 0 5 -1 9 19

. les phrases :

ex : twa penetre fɔrtəmãŋ dɑ̃ la dəmœ Rə

Locuteur féminin :
ou LF : 0 -7-6-3 -5-4 -3 -5 -1 -1 7 6

Dans ce cas, le tempo général est fait d'une succession de tempos croissants, c'est à dire de ralentissements, mais les différentes unités signalent leur commencement par une accélération relative instantanée, ce qui permet de bien délimiter début et fin de chacune d'elles.

2.1.2 Tempo à double variation (décroissant puis croissant) :

. les syntagmes :

ex : ɑ̃ dizãŋ bõnzur

LF :-6 7 -2 1 18

. les phrases

ex : sô mâŋto e taʃəte

LF : -1 4 2 0 -2-1 23

Ces tempos lorqu'ils appartiennent à des groupes, peuvent coexister dans la phrase avec des structures à tempo uniquement croissant.

Dans tous les cas, on constate que la "fin sémantique" d'un élément, quelle que soit sa place dans la hiérarchie de la phrase, coïncide avec un ralentissement net. Si une unité lexicale, c'est à dire un "mot" sans expansion, se termine par un tempo négatif, c'est que, pour le locuteur, son sens est ponc-

tuellement dépendant du suivant.

2.2 Etude d'un microcosme : les structures homophoniques :

Dans le cadre de cet article, nous n'envisagerons que les structures homophoniques qui permettent de rassembler en un contexte phonétique identique des éléments grammaticaux très divers aux relations syntaxiques tout aussi diverses. Ces structures offrent l'énorme avantage de montrer les corrélations qui se créent entre, parfois, la longueur des mots et le tempo, ou encore (et surtout) entre la syntaxe et ce dernier, en neutralisant les effets toujours possibles du contexte. C'est donc un domaine particulièrement privilégié assurant une fidèle représentation des phénomènes temporels, que l'on peut également étudier par ailleurs.

2.2.1 Les structures homophoniques à tempo croissant :

2.2.1.1 Appartenance grammaticale :

Elle est très diversifiée, que ce soit au niveau du mot lexical (substantif, verbe, adjectif, adverbe), du syntagme (nominal, verbal, prépositionnel au sens large) avec éventuellement leurs expansions respectives, et les différentes relations syntaxiques susceptibles d'exister entre ces unités. Toutefois, la nombre d'exemples de phrases totalement homophoniques est peu élevé.

2.2.1.2 Analyse formelle :

Il semble qu'il n'y ait pas de rapport entre les variations du tempo et la longueur des unités, sauf dans un cas, celui des monosyllabes lexicaux. Sur les 67 structures homophoniques, on relève 18 exemples de monosyllabes lexicaux, ou dissyllabes avec /ə/ final, constitués soit par des substantifs, soit des verbes. On constate que devant ces dernières les déterminants et pronoms personnels également monosyllabiques, sont beaucoup plus longs qu'à l'accoutumée, et possèdent un tempo positif.

Longueur moyenne des déterminants et pronoms personnels : 4 cs

Longueur moyenne des autres déterminants et pronoms personnels : -1.9 cs

Pour peu que le tempo de l'unité, par rapport à celuide groupe soit privigié par le locuteur. Les monosyllabes lexicaux ou dissyllabes à /ə/ final se comportent temporellement quant à eux, comme des syllabes finales de mots lexicaux.

Exemple :

	l	a	R	m	e		f	o	R	t	ə	-	l	a	R	m	ə		e		f	o	R	t
LM		-7			2		13				2			11			11		0		10			

	œ̃	s	i	ɲ	a	l	ə	m	ã	ŋ		-	œ̃	s	i	ɲ	a	l	ə	m	ã	ŋ
LF	-1	-1	-1	3	1	14							7	26		-4	0	32				

	õ	s	ãŋ	d	e	g	u	t	ə		-	õ	s	ãŋ	d	e	g	u	t	ə
LM	-4	2	4	13	14							1	9		0	10	14			

Il semble que généralement dans les structures homophoniques, à tempo croissant, les mots grammaticaux soient soudés au mot lexical suivant, sans qu'une accélération vienne souligner le début de celui-ci (77.4 % des cas). Les exceptions que l'on relève appartiennent à des mots lexicaux longs.

Exemple :

	œ̃	s	i	ɲ	a	l	ə	m	ã	ɲ̃	-	lə	p	ɛ	R	m	i	s	ɥ	o	n	ɛ	R
LM	2		-1		1		-3		11			0		-5		-2		2			25		

Lorsque deux mots grammaticaux précèdent l'élément lexical, le locuteur a la possibilité, souvent utilisée, d'intégrer temporellement le deuxième mot grammatical au tempo du groupe.

Exemple :

	s	y	R	œ̃	n	a	p	ə	r	õ	ɲ	-	s	y	r	y	n	ə	n	a	p	ə	R	õ	ɲ	d	ə
LF		2		-7		0		4		9				4		0	0			6	6		11		11		

Une unité lexicale monosyllabique se comporte comme une syllabe finale. Il est bien évident que dans ces conditions qu'un monosyllabe en fin de groupe ne pourra posséder de tempo démercatif.

Exemple :

	l	ə	f	ø	ɥ	a	ʒ	ə	v	ɛ	R
LF		-3		1		1		1		3	

2.2.1.3 Analyse quantitative

2.2.1.3.1. Les limites des unités temporelles

Pour comparer des faits semblables, parmi les unités à tempo simple croissant ayant défini temporellement leurs frontières, nous envisageons les débuts des unités temporelles au moment de l'accélération la plus grande, soit au niveau du mot grammatical initial, soit au niveau du mot lexical lorsqu'il n'est pas précédé de ce dernier, ou lorsqu'il se détache de lui par une accélération instantanée :

. début de l'unité : moyenne -2.5 cs
. fin de l'unité : moyenne 15.85 cs

soit un ralentissement en fin de mot de 17.3 cs par rapport à la syllabe initiale.

2.2.1.3.2. Tempo et lien syntaxique entre les unités

A l'opposé des structures attestant une rupture de continuité à l'intérieur d'un tempo croissant, celles qui manifestent un ralentissement progressif sans décrochement, ont été soulignées.

	Substantif + expansion	SN1 + SV	mot lexical sans lien syntaxique avec le suivant	V + SN2	V + G.prép.
Fin ↗	5.5	6.2	10.1	4.6	5.9
Début ↗	- 0.2	- 1.6	- 1.1	- 2.5	- 4.1
écart	- 5.7	- 7.8	- 11.2	- 7.1	- 10
Fin ↗	2.1	2		- 0.1	
Début ↗	7.5	8.5		1.3	
écart	+ 5.4	+ 6.5		+ 1.4	
Fin devant pause entre les 2 éléments cités	22.5 / P = 19	17.4 / P = 19.9	21.4 / P = 30.4	12.8 / P = 12.8	

TABLEAU 1 : Tempo et lien syntaxique entre les mots ou syntagmes
Tempo and syntactic links between various unities

Légende : SN1 syntagme nominal sujet
SN2 syntagme nominal complément d'objet
V verbe
Gprép groupe prépositionnel (au sens large)

* Note : il arrive parfois que les /ə/ situés à l'intérieur des mots se comportent de manière anarchique par rapport au tempo général de l'unité à laquelle ils appartiennent .

De ces résultats tirés de l'analyse des paires homophoniques, on peut conclure que la distance temporelle des différentes unités est d'autant plus brève que les unités sont syntaxiquement et sémantiquement plus liées : on constate en effet une cohésion plus forte entre les éléments interne du syntagme verbal (substantif et ses expansions, verbe et ses compléments d'objet) qu'entre le syntagme nominal sujet et le syntagme verbal, ou le verbe intransitif , ses adverbes et compléments circonstanciels, ou à plus forte raison entre des éléments sans rapport syntaxique individuel : les structures qui attestent un ralentissement progressif sans décrochement appartiennent de même aux unités les plus solidaires, et les valeurs prouvent l'opportunité de ces remarques. La longueur des pauses, plus que le tempo des syllabes devant celles-ci, vient confirmer encore nos résultats.

2.2.2 Les structures homophoniques à tempo double (décroissant puis croissant)

2.2.2.1 appartenance grammaticale et forme

L'appartenance grammaticale de ces unités ne présente aucune nouveauté par rapport à celle des unités précédemment étudiées. L'originalité se situe ailleurs, dans la forme même des constituants : en effet, tous les exemples de cette série concernent, soit au niveau du syntagme, de beaucoup le plus fréquent soit au niveau de la phrase, deux éléments lexicaux en rapport syntaxique et sémantique privilégié par le locuteur, avec les mots grammaticaux qui les accompagnent inévitablement, mais toujours l'un et l'autre sans expansion supplémentaire. Contrairement aux unités à tempo simple, les mots lexicaux non monosyllabiques se démarquent très souvent (93.3 % des cas) des mots grammaticaux par un ralentissement notable.

2.2.2.2 étude quantitative

A l'image de ce qui a été fait précédemment, on peut préciser la configuration temporelle de ces structures :

. élément grammatical monosyllabique initial -2.5 cs

Lorsque l'élément lexical qui suit est lui aussi monosyllabique ou dissyllabique avec /ə/ final, l'élément grammatical est également plus long (+ 1.5 cs)

. syllabe initiale du 1 er mot lexical 2.9 cs
. syllabe finale du 1 er mot lexical -1.6 cs
. syllabe initiale du 2 ème mot lexical -1.5 cs
. syllabe finale du 2 ème mot lexical 12.4 cs

La syllabe finale de l'ensemble est toujours très ralentie par rapport aux autres tempos syllabiques et dans des proportions beaucoup plus grandes que la syllabe initiale du 1 er mot lexical.

2.2.2.3 Tempo et lien syntaxique entre les unités

	Substantif + expansion	SN1 + SV	V + SN2	V + G. prép
Fin ↘	- 1.7	- 1.25	- 0.75	- 2.5
Début ↗	- 2.1	- 3	+ 0.75	+ 0.25
écart	- 0.4	- 1.75	+ 1.5	+ 2.75

TABLEAU 2 Tempo et lien syntaxique entre les mots ou syntagmes à tempo opposé (croissant ou décroissant)
Tempo and syntactic link between opposite tempo unities

Symétriquement par rapport au tableau précédent, on remarque que la distance temporelle la plus courte, réside entre les éléments internes des syntagmes (SN puis SV) fortement solidaires les uns des autres, puis entre le syntagme sujet et le syntagme verbal, puis entre le verbe intransitif et les adverbes ou groupes prépositionnels.

En ayant recours à ce schéma temporel double, le locuteur semble vouloir privilégier une relation syntaxique particulière et traduire une dépendance ponctuelle d'un élément d'une unité, comme le confirme par ailleurs, par rapport au tableau précédent, l'absence d'éléments sans lien syntaxique mutuel, et surtout celle des pauses.

CONCLUSION

On peut tirer de cette étude un certain nombre de conclusions :

1. L'utilisation du tempo dans les phrases, par les deux locuteurs, étudiés est similaire
2. Le tempo général est fait d'une succession de ralentissements après une accélération initiale. L'occurence d'une accélération en finale de mot actualise une dépendance ponctuelle d'un élément par rapport à celui qui le suit immédiatement, tandis que les ralentis internes de la phrase révèlent un sémantisme inachevé.
3. L'étude du tempo permet en outre de proposer des "patrons temporels" pouvant apporter une aide efficace à la segmentation en mots
4. Cette analyse nous permet en outre de préciser certains des rapports entre syntaxe et sémantique : il semble en effet que le cadre fixe des structures syntaxiques de la phrase, permette au locuteur d'opérer une série de choix successifs correspondant au sens particulier que celui-ci attribue à cette phrase. Il établit ainsi un lien entre une syntaxe aux sens virtuels et une sémantique aux sens effectifs.

- REFERENCES -

Geneviève CAELEN et G. MAURAND Fréquence, intensité et durée : étude comparative des fonctions dans les phrases énonciatives simples et étendues
7 èmes Journées d'Etude sur le Parole Nancy 1976 (G.A.L.F.)

Geneviève CAELEN et G. MAURAND Corrélations entre mélodie, intensité et durée dans la phrase française énonciative
8 èmes Journées d'Etude sur le Parole Aix-en-Provence 1977 (G.A.L.F.)

Geneviève CAELEN Etude prosodique des structures de la phrase énonciative
Actes du Séminaire C.E.R.F.I.A. Toulouse 1977

Geneviève CAELEN Structures prosodiques de la phrase énonciative simple et étendue
Thèse de 3 ème cycle (à paraitre) Toulouse 1978

Françoise EMERARD Synthèse par diphones et traitement de la prosodie
Thèse de 3 ème cycle Grenoble 1977

F. GOLDMANN-EISLER The distribution of pause durations in speech
Language and speech 4, 191-226 , 1961

F. GROSJEAN et A. DESCHAMPS Analyse des variables temporelles du français spontané
Phonetica 28, 191-226, 1973

V. LUCCI Etude phonostylistique du rythme et de la variabilité de la longueur en français parlé et français lu
Bulletin de l'Institut de Phonétiqe de Grenoble 2, 139-161 1973

V. LUCCI Rythme et longueur du message parlé. La conversation
Bulletin de l'Institut de Phonétique de Grenoble 3, 139-152 1974

M. SAINT-BONNET et J.L. BOE Les pauses et les groupes rythmiques : leur durée et distribution en fonction de la vitesse d'élocution
8 èmes Journées d'etude sur la Parole Aix-en-Provence 1977 (G.A.L.F.)

Jacqueline VAISSIERE Premiers essais d'utilisation de la durée pour la segmentation en mots dans un système de reconnaissance
8 ème Journées d'Etude sur la Parole Aix-en-Provence 1977 (G.A.L.F.)

9èmes JOURNEES D'ETUDE SUR LA PAROLE

LANNION 31 mai - 2 juin 1978

DETECTION AUTOMATIQUE DE SYLLABES ACCENTUEES EN NEERLANDAIS

A.C.M. Rietveld et L. Boves

Instituut voor Fonetiek
Katholieke Universiteit Nijmegen
Erasmuslaan 40
Nijmegen- Pays-Bas

.RESUME

Le but de nos procédures est d'arriver à la détection automatique de syllabes accentuées en néerlandais, comme premier pas vers une transcription automatique de l'intonation de cette langue.

A cet effet nous déterminons la fréquence fondamentale d'une phrase. Ensuite la courbe-Fo digitalisée est approximée par une série de lignes droites, d'après la méthode des moindres carrés. L'approximation d'une partie de la courbe est arrêtée dès que la variance autour de la ligne de régression ("l'erreur") a dépassé un maximum, choisi d'avance.
Les lignes droites, résultat de l'approximation, sont étiquettées et combinées par un programme qui fonctionne d'après quelques principes de la reconnaissance linguistique de patrons.

Le résultat de cette section-là, une concaténation de symboles accompagnés de quelques paramètres, est employé dans la section de détection proprement dite, joint à de l'information sur la structure syllabique et une mesure de l'amplitude des syllabes.

Un test préliminaire avec deux textes de courte durée ont donné des pourcentages de détection de 71% et 75%.

AUTOMATIC PROMINENCE-DETECTION IN THE DUTCH LANGUAGE

A.C.M. Rietveld and L. Boves

Instituut voor Fonetiek
Katholieke Universiteit Nijmegen
Erasmuslaan 40
Nijmegen - The Netherlands

SUMMARY

The procedures which will be described in this article aim at an automatic detection of prominence in Dutch. We regard this as a first step towards an automatic transcription of the intonation of the Dutch language. The physical correlates of prominence in this language are essentially 2 pitch movements, an early rise in the syllable or a late fall. Experiments with a vocoder ('t Hart, J., Collier, R., 1975) have shown that these prominence-lending movements can be stylized by straight lines; these lines are perceptually equivalent to the original movements.

In the detection-procedures we have to find those fragments of the Fo-curve which can be regarded as realisations of the prominence-lending movements mentioned above. Therefore, having measured the Fo-variations with an analog pitchmeter, the resulting curve is converted into a digital signal for further processing. First a correction and smoothing program smooth the rather irregular curve. After these programs have been finished an approximation procedure is carried out; this procedure transforms the curve into a series of straight lines, obtained by applying the Least-Square criterium and also an "error" -criterium which stops the approximation with the aid of a single line if the error of that approximation has a greater value than the "reference-error".

The approximation procedure provides the input of a labelling program. This programlabels the line-segments in accordance with the values of some parameters (rise,fall,length etc) and tries to combine similar lines into lines which are then given the same label. This program is partly based on the principles of linguistic pattern recognition. The labels of the line-segments and their parameters are used in the final phase: the detection of prominence.

In this phase we make use of the amplitude of the signal and the segmentation of the utterance into syllabic units which is carried out by the syllabledetector. Only those syllables with an amplitude above the median of the amplitudes of the individual syllables are regarded as "candidates" for prominence. If a pitchmovement is also present in the syllable which can be considered as characteristic of the prominence-lending movements (and fulfils certain requirements concerning Fo-movements and place in the syllable) the syllable in question is classified as "prominent".

Preliminary tests with two short texts resulted in detectionscores of 71% and 75%.

INTRODUCTION

La recherche perceptive sur la structure intonative du néerlandais a montré que, pour la perception, il n'y a qu'un nombre limité de mouvements de hauteur pertinents ('t HART, J. et COLLIER, R., 1975). Les mouvements peuvent être combinés pour former un certain nombre de contours intonatifs qui, à leur tour, constituent les patrons intonatifs du néerlandais. La manipulation de la courbe mélodique à l'aide d'un vocoder (WILLEMS, L.F., 1966) a mis en évidence qu'une série de mouvements linéaires montants et descendants entre deux niveaux est une bonne approximation de la courbe mélodique originale; une telle approximation n'a presque pas de conséquences perceptives. C'est pourquoi les mouvements de hauteur sont représentés schématiquement par des lignes droites et symbolisés par une douzaine de symboles: 1,2,3,0,Ø,A,B etc.. C'est avec ces symboles-là que les auteurs cités plus haut sont capables de transcrire l'intonation d'un très grand pourcentage des phrases qu'on leur présente. L'analyse perceptive a même montré que la combinaison des mouvements intonatifs est réglée par une"grammaire intonative".

ANALYSE DU PROBLEME

Les procédures que nous allons décrire ici ont été développées dans le but d'arriver à une détection automatique des syllabes accentuées en néerlandais et pour cela, à une détection automatique des mouvements mélodiques symbolisés avec "1" et "A" d'après le système de 't HART et al.. Ce sont des mouvements de hauteur montants et descendants - symbolisés respectivement par "1" et "A" - qui accordent le trait "proéminent" aux syllabes dans lesquelles ils se trouvent.

Nous considérons la détection des "1" et "A" comme le premier pas vers une transcription automatique de l'intonation en néerlandais. Les mouvements "1" et "A" constituent les "clefs de voûte" de la transcription de l'intonation, qu'il s'agisse d'une transcription auditive ou d'une transcription automatique.
Avant de continuer nous mettrons au point quelques termes qui seront employés dans ce qui suit:

Courbe mélodique: la courbe qui représente les variations de l'inverse de la période fondamentale
L'intonation : le phénomène perceptif qui a comme corrélat physique la fréquence fondamentale.

La courbe mélodique a d'ordinaire une forme très irrégulière et ne ressemble souvent que très vaguement à la représentation schématique de la transcription auditive: une concaténation de lignes droites. Cette irrégularité et cette déviation du "schéma" sont dues à plusieurs facteurs:

1. La concaténation de lignes droites ne forme qu'un modèle de la courbe mélodique: elle est un schéma perceptif
2. Les voyelles et les consonnes voisées ont des fondamentales intrinsèques (cf. LEHISTE, I., 1970). Les différences entre les Fo des sons divers peuvent s'élever à 15 Hz ou plus.
3. Il y a souvent des mouvements brusques à la jointure de consonnes et voyelles.
4. La périodicité du signal n'est pas toujuors définie.

PROCEDURE

Si nous voulons trouver les fragments de la courbe mélodique qui peuvent être considérés comme les réalisations de mouvements linéaires ayant les traits caractéristiques des mouvements "1" et "A", nous devons procéder à des transformations du signal pour le rendre plus apte à l'extraction des dits mouvements.

Dans le schéma suivant nous indiquons les opérations successives utilisées pour obtenir de l'information qui est plus propre aux procédures de reconnaissance:

1. analyseur de mélodie
   A/D conversion
2. programme de correction
3. filtrage passe-bas
4. approximation par lignes droites
5. étiquettage et réécriture

l'analyseur de mélodie est un instrument analogue dont le fonctionnement est basé sur la détection de crêtes. Pour une description plus détaillée nous faisons référence à van ROSSUM, N. et BOVES, L. (1978) et à l'appendice A.

La courbe mélodique sortant de l'analyseur de mélodie est convertie en signal digital. Ensuite un programme entre en fonction qui a pour but de "corriger" la courbe de Fo telle qu'elle est fournie par l'analyseur. Le programme rejette les valeurs de la fondamentale qui sont clairement des ratés et les remplace par des valeurs plus probables, si possible. Le processus de "rabotage" de la courbe mélodique finit par un filtrage passe-bas, avec une fréquence de coupure de 100 Hz (8e degré).

Le résultat de ces opérations est une courbe plus égalisée, il est vrai; cependant la forme en reste trop irrégulière pour nos procédures de reconnaissance. C'est pourquoi dans la phase suivante (4) une approximation de la courbe par une série de lignes droites est introduite. L'approximation par des lignes droites a été choisie pour plusieurs raisons:

1. La stylisation d'une courbe mélodique par des lignes droites ne semble pas avoir des conséquences perceptives considérables (ROMPORTL,M., 1972).
2. La représentation schématique de la transcription de 't HART et al. se fait largement à l'aide de lignes droites.
3. La ligne droite constitue une unité d'information (anglais: "primitive") qui est convenable aux procédures de reconnaissance qui suivent.

Pour l'approximation par des lignes droites nous faisons usage du critère des moindres carrés (KLOKER,D., 1977). Il sera clair que le fragment de la courbe Fo qui, à première vue, peut être bien représenté par une ligne droite, ne sera pas "parfaitement" approximé par cette ligne: la ligne droite n'est qu'un modèle. L'approximation d'un segment réel par ce modèle amènera une faute: "l'erreur". Tant que l'erreur n'est pas trop grande, nous acceptons l'approximation obtenue. L'erreur se calcule comme suit:

$$s_e^2 = \frac{\Sigma (Y_i' - Y_i)^2}{n - 2}$$

Y_i = les valeurs Fo observées
Y_i' = les valeurs correspondantes de l'approximation
n = le nombre de périodes de l'approximation

L'erreur que nous avons trouvée dans l'approximation d'un segment de la courbe que nous jugeons subjectivement "plus ou moins droit" est appelée "l'erreur standard": la valeur de cette erreur-standard est adaptée à la médiane de la fondamentale: elle varie de 2.0 à 2.6 pour n=11.
L'approximation d'un segment de la courbe mélodique par une ligne droite est acceptable tant que l'erreur qui a été calculée pour une approximation n'est pas plus grande que l'erreur-standard. Avec le F-test nous pouvons comparer les deux err-

eurs et déterminer s'il y a raison de rejeter l'hypothèse nulle qui dit que les deux erreurs sont égales (seuil de rejet: $p < 0.025$).

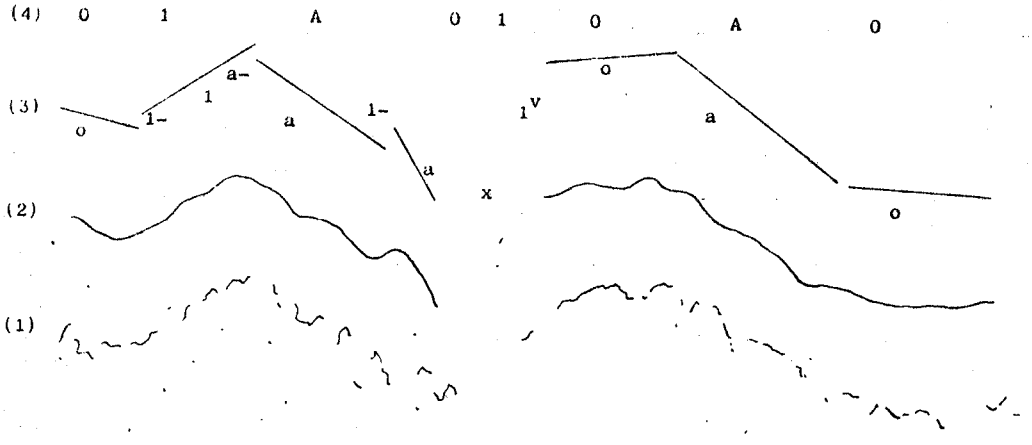
La procédure d'approximation commence par 3 valeurs Fo; puis l'erreur est calculée et enfin le F-test est appliqué. S'il n'y a pas de différence significative entre les 2 erreurs, la période suivante est ajoutée et on recommence la procédure la procédure. On continue à ajouter des valeurs de la fondamentale tant que la différence entre les 2 erreurs n'est pas significative. Dès qu'une différence significative se produit, on arrête l'approximation et on garde les paramètres de l'avant-dernière approximation. Ensuite l'approximation suivante est commencée. Après que les phases 1,2,3 et 4 ont été accomplies, une série de lignes droites a été obtenue, dont nous connaissons: 1. Début, 2. Fin, 3. Valeur init. de la fondamentale, 4. Valeur finale de la Fo.

Dans la phase suivante (5) nous procédons à l'étiquettage des lignes droites en leur appliquant des symboles en fonction du caractère des mouvements de la fondamentale:

1	:	montée
a	:	chute
1^v	:	montée virtuelle
a^v	:	chute virtuelle
1-	:	petite montée virt.
a-	:	petite chute virt.
o	:	niveau (pas de mouvement)
x	:	segment indéfini

Une procédure de "réécriture" succède à la phase d'étiquettage. A la base de cette procédure se trouve l'hypothèse que la concaténation de symboles assez divers comme nous les avons énumérés plus haut, peut être la "réalisation" d'un contour plus "abstrait". Cela veut dire qu'un contour 01A (niveau, montée, chute) tel qu'on le trouve dans la transcription de 't HART peut se manifester à la sortie de la phase d'étiquettage comme la concaténation:

o 1- 1 a- a 1- a (Fig. 1)



"Heleen wil die kleren meenemen"

Fig.1 (1) courbe mélodique
(2) courbe mélod. corrigée
(3) lignes droites étiquettées
(4) symboles de transcription du contour "sous-jacent"

(1) raw pitch-curve
(2) smoothed pitch-curve
(3) labelled straight lines
(4) transcription symbols of the "underlying" pitch-contour

On pourrait supposer que de telles réalisations sont générées par un nombre de règles de réécriture comme p.e.:

S → IOA A → Aa,a
I → I1,1 O → Oo,o

Dans nos procédures de reconnaissance nous n'avons pas à générer les concaténations de symboles, mais nous devons être capables de déterminer si une séquence telle que 1 a- 1 1v a été générée par des règles dont I → I1,1 fait partie.

En fait, de cette façon-là on essaie d'imiter le phonéticien qui inspecte des courbes mélodiques; il ne tient pas compte de petites déviations de la ligne "générale" mais il s'intéresse à la courbe globale. Nos procédures peuvent être mises au nombre des "techniques linguistiques de reconnaissance de patrons". En effet, les règles de réécriture hypothétiques ressemblent aux règles que nous connaissons de la grammaire générative.

L'information provenant du programme d'étiquettage et de réécriture consiste dans une série de symboles; chacun de ceux-ci est accompagné de 5 paramètres: temps init., temps final, Fo-init., Fo-finale et une variable qui indique si le segment, représenté par le symbole, se trouve devant un segment non-voisé.
Les symboles utilisés sont I, A et O. Il ne faut pas les confondre avec les symboles utilisés dans les transcriptions de 't HART et al.. Au contraire: une concaténation comme I O A n'est qu'une première hypothèse sur la forme pertinente du patron intonatif du fragment en question. Cette hypothèse sera utilisée dans la dernière phase du processus: la procédure de détection de syllabes accentuées.

Le schéma de la Fig. 2 présente l'ensemble des procédures. La procédure finale (la détection d'accents proprement dite) fait usage de 4 types d'information:

1. la courbe mélodique
2. l'intensité
3. la structure syllabique

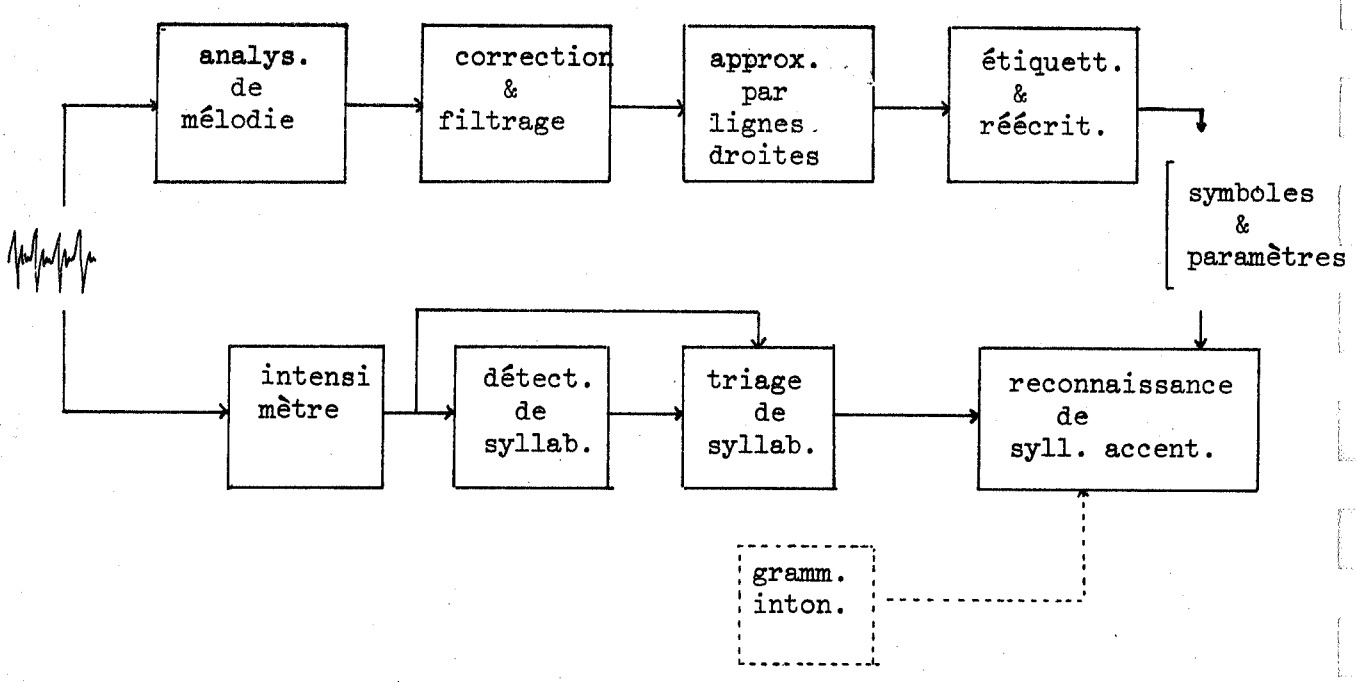


Fig.2 Schéma des procédures de détection

Block diagram of the prominence-detection procedures

Dans les pages précedentes nous avons parlé de la première source d'information. Restent les trois autres sources. Bien que l'intensité ne joue pas le rôle dans l'accentuation qu'on lui attribue intuitivement - en néerlandais son rôle perceptif semble être négligeable (v.KATWIJK, A.F.V., 1974) - nous avons constaté qu'elle peut nous donner de l'information sur "l'accentuabilité" des syllabes (cf. RIETVELD, A.C.M., 1975). Des expériences pilotes ont montré que l'amplitude des syllabes accentuées est souvent supérieure à l'amplitude moyenne. Ce n'est pas étonnant: un mouvement ascendant de la fondamentale est souvent accompagné d'un accroissement de l'intensité (BOE, L-J., RAKOTOFIRINGA, H. 1975).

La structure syllabique d'une phrase joue un rôle très important, pour plusieurs raisons:

a. les symboles de transcription de 't HART et al. se rapportent aux différentes syllabes de la phrase.
b. la place des mouvements par rapport au début ou à la fin de la syllabe peut aider à distinguer les mouvements les uns des autres.

Avant que la détection de syllabes ait lieu, le signal dont on mesure l'amplitude est filtré par un filtre passe-bande de 300 à 3000 Hz (cf. MERMELSTEIN, P., 1975). De cette façon-là on limite l'amplitude de consonnes comme le /s/ qui ont de l'énergie surtout dans les hautes fréquences. L'intensité mesurée par l'intensimètre (temps d'intégration: 10 ms.) est convertie en signal digital avec une fréquence d'échantillonnage de 500 Hz; pouvoir séparateur: 12 bits.
Ensuite la détection de limites de syllabes se fait sur ce signal.

D'abord le début et la fin de la courbe d'amplitude sont déterminés en cherchant une valeur qui dépasse un certain seuil. Puis on fixe les valeurs suivantes:
seuil de durée (SD): 250 ms.
seuil de différence d'amplitude (SA): 300
Ensuite le programme cherche le premier maximum et le premier minimum dans la courbe. Si le délai entre la limite de syllabe à gauche (ou le début de la phrase) et la place du minimum dépasse SD et la différence entre le maximum et le minimum est plus grande que SA, le minimum est accepté comme limite de syllabe et on continue la procédure. Cependant s'il n'y a pas de minimum qui remplisse ces conditions, on donne une valeur de 350 ms. à SD et une valeur de 150 à SA; puis on recommence la procédure. Si l'on ne trouve toujours aucune limite, on attribue 500 à SD et 50 à SA. (si nécessaire SA peut être égal à 0). Après que toute la phrase a été traitée, on examine les durées des syllabes. Les syllabes ayant une durée de moins de 50 ms. sont jointes à leurs voisines de droites pourvu que le minimum qui détermine le début de ces syllabes ait un niveau plus bas que le minimum qui en détermine la fin: sinon on les relie à la voisine de gauche.

Nous finissons par la description du programme de détection de syllabes proprement dit.

D'abord l'amplitude maximum de chaque syllabe est déterminée. Puis on calcule la médiane de ces amplitudes. Les syllabes ayant une amplitude supérieure à la médiane sont considérées comme des "candidats d'accentuation".

Puis on cherche des "I" et des "A" qui se trouvent dans ces syllabes et qui remolissent les conditions suivantes:

a. Pour "I" et "A": la fondamentale doit monter de plus de 20% en 100 ms. la montée totale doit dépasser 15% de la valeur initiale.
b. La montée doit commencer avant que 66% de la durée de la syllabe se soient écoulés.
c. Le mouvement descendant ne doit pas commencer avant que 50% de la durée de la syllabe se soient écoulés.
d. Si le mouvement total est inférieur à 30% de la valeur initiale (en Hz), le mouvement doit franchir une ligne imaginaire formée par la moyenne de la fondamentale. (avec une correction pour la fin de la phrase: on sait que l'intonation globale descend doucement vers la fin de la phrase)

QUELQUES RESULTATS PRELIMINAIRES

Vu que nos procédures sont en cours d'élaboration (fin Février) nous n'avons pas encore eu le temps de mettre à l'épreuve l'ensemble de nos procédures. Toutefois, deux textes de courte durée ont été analysées dans le but d'en extraire automatiquement les syllabes accentuées. Les syllabes accentuées sont celles qui avaient été désignées comme "accentuées" par la majorité d'une dizaine de sujets.

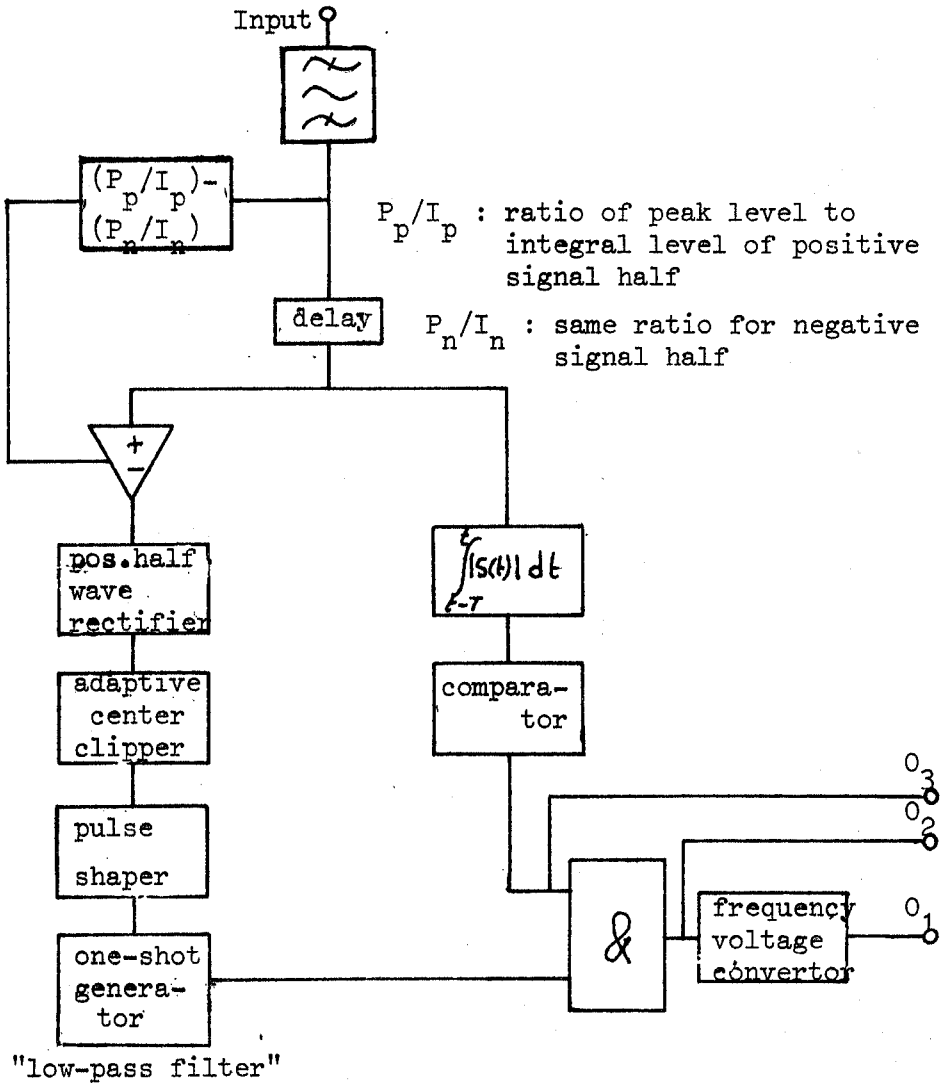
1er texte: lu par un des auteurs; il comprenait 33 accents sur 200 syllabes.
pourcentage de détection: 72%.
2ème texte: un bulletin météorologique radiodiffusé; celui-ci comprenait 24 accents sur 124 syllabes.
pourcentage de détection: 75%.

DISCUSSION

Le système est loin d'être prêt. On peut encore améliorer les systèmes de règles dans les programmes de réécriture et de détection de syllabes accentuées. En plus il faut trouver les moyens de détecter les accents qui ne sont pas accompagnés de variations de hauteur, mais qui sont surtout causés par des valeurs "inattendues" des paramètres intensité et durée.

REFERENCES

BOE, L-J., RAKOTOFIRINGA, H., 1975, A statistical Analysis of Laryngeal Frequency: its Relationship to Intensity Level and Duration. - Language and Speech 18, pp. 1-33.
FU, K.S., 1974 , Syntactic Methods in Pattern Recognition. Acad. Press, New York.
't HART, J., COLLIER, R., 1975, Integrating Different Levels of Intonation Analysis. - Journal of Phonetics 3, pp. 235-255.
KATWIJK, A.F.V., 1974, Accentuation in Dutch. Diss., Utrecht.
KLOKER, D., 1976, A Technique for the Automatic Location and Description of Pitch Contours. - IEEE int. Conf. on Acoustics, Speech and Signal Processing, pp. 55-59.
LEHISTE, I., 1970, Suprasegmentals. Cambridge Mass. M.I.T. Press.
MERMELSTEIN, P., 1975, Automatic Segmentation of Speech into Syllabic Units. - JASA 58, pp.880-883.
RIETVELD, A.C.M., 1975, Wordboundaries in the French Language. - Proc. of the Inst. of Phonetics Nijmegen no. 1. pp. 21-40.
ROMPORTL, M., 1972, Zur Synthese der Intonation. Acta Universitatis Carolinae, Philologica 2, Phonetica Pragensia IV.
ROSSUM, N. van, BOVES, L., 1978, an Analog Pitch-Period Extractor. - Proc. of the Inst. of Phonetics Nijmegen no.2. pp. 1-17.
WILLEMS, L.F., 1966, the Intonator. -IPO Ann. Prog. Rep. 1. pp. 123-125.



O_1: Analog pitch output

O_2: TTL pitch pulse output

O_3: Voiced/Unvoiced output (TTL compatible)

Block diagram of the IFN Pitch-period detector

APPENDICE A

9èmes JOURNEES D'ETUDE SUR LA PAROLE

LANNION 31 mai - 2 juin 1978

ETUDE ET UTILISATION DES INDICES ACOUSTIQUES ET DES TRAITS POUR LA SEGMENTATION ET LA RECONNAISSANCE PHONEMIQUE DE LA PAROLE

Claude SANCHEZ, Gérard MESSENET, Jean-Paul HATON
Equipe de Traitement du Signal et Reconnaissance des Formes
Centre de Recherche en Informatique de Nancy.

.RESUME

Cet article présente le niveau acoustico-phonétique de traitement du système MYRTILLE pour la compréhension du discours parlé continu, et particulièrement le rôle joué par les indices acoustiques et les traits dans les différents processus de ce niveau.

Le système en cours de réalisation permet de définir des méthodes et des stratégies d'utilisation des paramètres extraits du signal vocal pour l'identification des segments phonémiques, préalablement segmentés. Dans ce système le traitement est réalisé par des algorithmes organisés selon une structure arborescente modifiable. Chaque algorithme se fonde sur certains traits et indices pour conditionner la suite du traitement.

L'adjonction ou la suppression d'algorithmes, l'étude des stratégies d'utilisation d'un indice donné sont des manipulations aisées. Ce système est ainsi un outil expérimental très intéressant et c'est de plus un méta-système capable d'engendrer la structure optimale d'utilisation d'un ensemble de paramètres donné.

Après avoir rappelé le principe de la segmentation phonémique utilisée, nous présentons l'architecture du système ainsi que la structure et les principes de deux types d'algorithmes.

Experimental study and use of acoustic features for phoneme segmentation and identification in connected speech

· SUMMARY

This paper deals with the acoustic-phonetic processing level of MYRTILLE Speech Understanding system presently under development in our Laboratory. More precisely, the phonemic segmentation and the use of acoustic features in phoneme identification are described.

This system directly works on the digitized speech wave, which makes it possible to get a better accuracy on the parameters. The segmentation process is based on the computation of a distance between successive, 300 milliseconds portions of signal by the use of acoustic parameters such as energy, zero-crossings rate, length of the curve, etc...

The phoneme identification system has a tree structure in which the nodes are algorithms. These algorithms use one or several acoustic features in order to weight their answers. These answers can be phoneme classes and/or other algorithms. This structure is very convenient for studying the accuracy of various parameters in phoneme recognition and for implementing various recognition strategies. It appears like a meta-system which is able to generate an optimal recognition system for a given subset of acoustic parameters.

This system is presently under development and results will be presented during the Poster-session of the Seminar.

INTRODUCTION

Le système MYRTILLE II en cours de réalisation dans notre Laboratoire est conçu pour la compréhension de la parole continue dans le cas de langages naturels ou pseudonaturels. Nous présentons dans cet article l'étage acoustico-phonétique de traitement chargé de fournir la transcription d'une phrase sous forme d'un treillis de phonèmes, et plus particulièrement la façon dont les indices acoustiques sont utilisés à ce niveau.

Ce système travaille directement sur le signal de parole numérisé, ce qui permet une plus grande souplesse et plus grande variété des traitements effectués. Ce choix correspond à la philosophie générale du système que nous allons décrire et qui se présente comme un système expérimental destiné à tester l'efficacité en reconnaissance des paramètres extraits du signal vocal et de leurs diverses stratégies d'utilisation.

L'identification des unités phonétiques dans la parole continue est précédée d'une phase de segmentation utilisant un certain nombre d'indices déduits du signal acoustique et de traits (intensité, voisement, taux de passages par zéro, longueur curviligne du signal). Le principe de cette segmentation est décrit dans une première partie. Nous présentons ensuite le système d'identification de phonèmes qui possède une structure générale arborescente. Les algorithmes composant cette arborescence utilisent un ou plusieurs traits ou indices pour pondérer les réponses qu'ils fournissent et conditionner ainsi la suite du processus. L'architecture utilisée permet l'adjonction ou la suppression d'algorithmes (donc la prise en compte de divers paramètres), ainsi que le choix de différentes stratégies d'utilisation des paramètres.

Pour un ensemble de traits et d'indices donné, notre système permet d'obtenir la structure de reconnaissance correspondant à une utilisation optimale de ceux-ci. Cette structure est fondamentalement destinée à être implantée sous une forme répartie (multiprocesseurs) ; néanmoins, la version actuelle fonctionne sur un ordinateur monoprocesseur SEMS Mitra 125.

SEGMENTATION

Nous allons présenter dans ce qui suit les différentes approches que nous avons envisagées pour réaliser la segmentation du discours parlé. Les méthodes seront différentes, bien sûr, suivant le but à atteindre c'est-à-dire suivant l'entité que l'on désire localiser. Nous n'aborderons pas ici la segmentation en mots qui résultera d'une analyse lexicale sur la chaîne ou le treillis de phonèmes fourni par le module de reconnaissance phonémique.

Nous ne parlerons pas non plus de la segmentation en groupes de sens qui, elle, utilisera des indices prosodiques (pitch, phonèmes de longue durée, phonèmes de grande intensité), mais nous nous intéresserons seulement à la segmentation phonémique du signal temporel, segmentation dont le résultat sera utilisé par le module de reconnaissance des phonèmes.

Pour réaliser cette segmentation, la méthode envisagée a été basée sur un calcul de paramètres acoustiques sur des fenêtres successives se recouvrant partiellement. La notion de recouvrement ayant pour but d'opérer un lissage sur les paramètres calculés. Pour décider de poser une marque de segmentation, dans une fenêtre donnée, il faut donc trouver une méthode de comparaison des valeurs de ces paramètres d'une fenêtre à l'autre, ce qui revient à calculer une distance entre fenêtres de façon à répartir ces fenêtres en deux classes : fenêtres non transitoires si elles sont extraites d'un même phonème, fenêtres transitoires si elles contiennent le passage d'un phonème à un autre.

Les paramètres envisagés l'ont été non pas pour leur signification physique, même s'ils en ont une, mais plus pour se placer dans l'optique de reconnaissance des formes. Il est en effet intéressant de constater qu'une segmenta-

tion manuelle correcte peut être obtenue à partir du signal temporel visualisé, en utilisant des indices acoustiques présents dans ce signal.

Ces paramètres sont les suivants :

- le nombre de passages par zéro, ou plus exactement, le nombre de traversées moyen par seconde de la bande $(-\varepsilon , +\varepsilon)$.
- l'énergie absolue moyenne
- l'énergie quadratique moyenne
- le nombre de passages par zéro de la dérivée du signal temporel
- la longueur curviligne moyenne, scindée en longueur moyenne des portions de signal à valeurs positives et en longueur moyenne des portions à valeurs négatives, ce qui permet d'introduire un autre paramètre qui est le rapport de ces 2 dernières.

Le principe de la segmentation ayant été exposé ci-dessus, il reste à définir le choix de la distance permettant de comparer une fenêtre aux précédentes. Dans un premier temps, nous avons recherché une distance qui ait une justification mathématique. Une transition étant caractérisée par une variation plus ou moins grande des différents paramètres, l'idée est de comparer cette variation entre deux fenêtres consécutives par rapport à la variation entre les deux fenêtres précédentes. Cette comparaison est faite en calculant le rapport des deux variations. Si $p_{j-2}(i)$, $p_{j-1}(i)$, $p_j(i)$ sont les valeurs prises par le ième paramètre sur 3 fenêtres successives, la distance de variation de ce paramètre i , sera $d_j(i) = \dfrac{p_j(i) - p_{j-1}(i)}{p_{j-1}(i) - p_{j-2}(i)}$ et la distance ramenée sur tous les paramètres que l'on calculera sera $d_j = \sum_{i=1}^{i=p} d_j(i)$. Si la fenêtre j est non transitoire, les nombres $p_{j-2}(i)$, $p_{j-1}(i)$ et $p_j(i)$ seront proches et $d_j(i)$ sera proche de 1 ; par contre si la fenêtre est transitoire $p_{j-2}(i)$ et $p_{j-1}(i)$ seulement seront voisins, ainsi chaque $d_j(i)$ et donc d_j seront grands par rapport à 1 . La décision fenêtre transitoire ou non transitoire se fait donc sur comparaison de la distance d_j à un seuil fixé. Les résultats obtenus jusqu'à présent montrent que pour une longueur de fenêtre de 30ms (300 échantillons à 10 000 hertz) et une translation de 20ms, il faut prendre un seuil voisin de 0,8.

Cette distance a plusieurs inconvénients auxquels on peut plus ou moins remédier : tout d'abord des inconvénients dus à la nature du signal à analyser. Si le signal était parfaitement périodique, il faudrait par exemple, pour que les nombres $p_i(j)$ soient voisins sur des fenêtres consécutives, que la longueur de fenêtre ou la translation de ces fenêtres soit un nombre entier de périodes. Pour y remédier il faut donc lier la translation à la valeur du fondamental ou au nombre de passages par zéro. De plus certains paramètres varient parfois beaucoup, même sur des fenêtres non transitoires, par exemple les perturbations d'une liquide sur la voyelle suivante introduiront une variation du nombre d'extrema au cours de la voyelle.

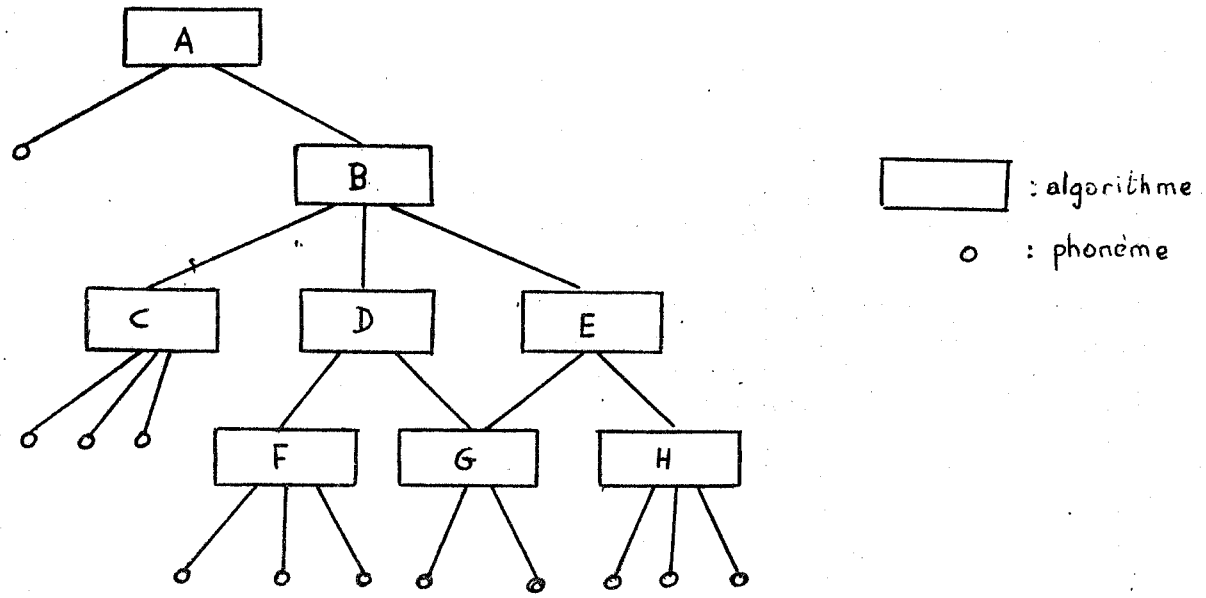
Enfin un autre inconvénient provient du fait que si l'on recherche une méthode de segmentation rapide, il serait préférable de n'évaluer qu'un minimum d'indices, alors que pour segmenter certains cas ambigus, il pourrait être intéressant d'en évaluer plus. Nous envisageons donc actuellement un type de distance plus heuristique ou l'évaluation des paramètres, ainsi que le calcul de la distance se ferait de façon arborescente. Les indices les plus discriminants étant calculés les premiers (par exemple : énergie, nombre de passages par zéro),

en fonction de cette évaluation il sera la plupart du temps, possible de décider si la fenêtre est transitoire ou non. C'est en particulier le cas pour la segmentation parole-non parole. Ce n'est que lorsque l'on ne pourra pas décider de l'existence d'une transition que l'on calculera d'autres paramètres. Le premier paramètre étant calculé, le choix du ième paramètre à calculer éventuellement dépendra donc des (i-1)èmes. La distance à envisager pour chaque paramètre peut être du type $d_j(i)$ comparée à un seuil $s(i)$ propre à chaque paramètre.

Quelle que soit la distance choisie, de toute façon il est peu probable que l'on arrive à segmenter le signal temporel de façon sûre, c'est pour cela que nous avons retenu la structure de treillis de segments comme résultat de la segmentation, les frontières de phonèmes étant affectées d'un score proportionnel à la distance calculée précédemment.

RECONNAISSANCE PHONEMIQUE

Le niveau acoustico-phonétique de Myrtille permet de définir des systèmes d'identification de phonèmes ayant une structure de ce type :



où les rectangles représentent des algorithmes de discrimination entre phonèmes faisant appel à certains indices acoustiques ou traits de la parole.

Chaque algorithme propose des phonèmes comme solution ou bien demande l'exécution d'autres algorithmes. L'algorithme joint à ses réponses une pondération. Les poids ainsi fournis permettent le calcul de scores qui guident le cheminement à travers la structure.

Afin d'éviter de s'engager trop souvent dans de mauvaises voies et afin de trouver rapidement la bonne réponse dans les cas non ambigus, un système d'identification de phonèmes basé sur ce modèle doit suivre un minimum de règles élémentaires.

Les algorithmes à décision binaire, à réponses franches ou faisant appel à peu de calculs doivent être utilisés le plus tôt possible. Des algorithmes à réponses moins sûres ou bien nécessitant des calculs complexes et longs doivent être exécutés le moins souvent possible à tort et donc se trouver à l'extrémité d'un chemin dans la structure.

Nous présentons ci-dessous deux exemples extrêmes d'algorithmes d'identification.

(i) Si l'on observe de façon grossière le nombre de passages par zéro du signal de parole par unité de temps on voit que le phonème |s| se distingue très nettement de tous les autres avec un nombre de passages par zéro très élevé. Le nombre de passages par zéro étant un paramètre déjà calculé pour la segmentation il est disponible. La reconnaissance des |s| peut donc se faire sur un simple test. Ce test étant rapide, simple et efficace, il sera fait, dès l'entrée dans le système d'identification, par le premier algorithme.

A une étape plus avancée dans le processus d'identification, une observation plus fine du nombre de passages par zéro pourra, conjointement à d'autres paramètres, permettre une discrimination entre d'autres phonèmes.

(ii) Pour identifier une voyelle nous considérons ses trois premiers formants (F1, F2, F3).

Pour déterminer ces trois formants nous calculons des coefficients de prédiction linéaire sur la partie stable de la voyelle par la méthode d'autocorrélation. Une transformée de Fourrier sur ces coefficients fournit le spectre de la fonction de transfert du filtre inverse du conduit vocal. Nous prenons alors les fréquences des maxima de ce spectre comme valeurs de formants.

Lors d'un apprentissage préalable sur un corpus, les formants sont calculés sur toutes occurences de chaque voyelle. Pour chaque voyelle la moyenne et la variance des formants est mémorisée. Chaque voyelle est alors considérée comme une classe d'individus, caractérisé par son centre de gravité dans l'espace (F1, F2, F3) et par sa dispersion selon les axes.

L'identification d'une voyelle se ramène alors au calcul de sa distance aux différentes classes. La distance utilisée est une distance euclidienne pondérée. Le coefficient affecté à chaque formant est différent suivant la voyelle à laquelle est comparé le segment inconnu. Ces coefficients sont déterminés en deux étapes : un premier coefficient réduit la dispersion à l'intérieur des classes, un deuxième favorise les formants les plus discriminants pour chaque voyelle.

Les distances calculées, on en déduit les proximités en les inversant et des poids en les normalisant.

L'algorithme d'identification de voyelles fournit alors la liste des voyelles accompagnées des poids ainsi calculés.

Les résultats expérimentaux du système seront présentés lors de la session d'exposition des JEP.

CONCLUSION

Le système de reconnaissance acoustico-phonétique que nous venons de décrire brièvement a été conçu comme un système expérimental d'étude du rôle des indices et des traits utilisables en reconnaissance et de leurs stratégies d'utilisation. Il n'est donc pas performant en lui-même mais permettra d'engendrer un système optimal -tant en pourcentage de reconnaissance qu'en temps de calcul- pour un ensemble de paramètres donné.

9èmes JOURNEES D'ETUDE SUR LA PAROLE

LANNION 31 mai - 2 juin 1978

LE MODELE LINEAIRE DE PRODUCTION DE LA PAROLE UTILISE POUR LA RECONNAISSANCE AUTOMATIQUE DE VOYELLES NON NASALISEES DE LA LANGUE FRANCAISE

Y. RENE de COTRET

Laboratoire de Physique Générale
Université de Liège,
B-4000 SART TILMAN par Liège 1, Belgique.

.RESUME

Les paramètres du modèle linéaire de production de la parole, coefficients prédicteurs et coefficients de réflexion, permettent d'obtenir pour la reconnaissance de voyelles de la langue française, prononcées par 16 locuteurs masculins et féminins, des taux élevés de reconnaissance, pour autant que l'on utilise une fonction discriminante appropriée. Il est en outre possible d'obtenir une augmentation des taux de reconnaissance si, au cours de la phase d'apprentissage, on distingue les locuteurs masculins des locuteurs féminins.

Les travaux réalisés portent sur 2875 occurences de voyelles, chacune étant prononcée dix-huit fois par 8 hommes et 8 femmes. Ces voyelles ont été prononcées soit isolément, soit à l'intérieur de mots polysyllabiques, soit encore à l'intérieur d'une phrase. Le choix de différentes fonctions discriminantes et de conditions particulières quant à l'ordre du système, aux fichiers traités et aux classes de locuteurs, nous a conduit à 44 expériences de reconnaissance.

LINEAR PREDICTION OF SPEECH APPLIED TO AUTOMATIC RECOGNITION OF NON NASALISED FRENCH VOWELS

Y. RENE de COTRET

Laboratoire de Physique Générale,
Université de Liège,
B-4000 SART TILMAN par Liège 1, Belgique

.SUMMARY

We have chosen a representation of ten French vowels derived from linear prediction analysis as presented by ITAKURA and SAITO. Three types of parameters were retained: predictor coefficients, reflection coefficients and autocorrelation coefficients. The latters preserve all the information retained about the signal, the previous ones add what comes from the linear prediction model.

Parameters obtained from linear prediction are of great interest. A small number represents in the time domain all of the speech spectrum envelope and, mathematically, they fo not depend upon the lenght of the acoustic tube modeling of the vocal tract they represent. Moreover as they are related to the shape this acoustic tube modeling, we may hope to distinguish male from female speakers since women have shorter pharynges in relation to their oral cavities. But for this same reason these parameters are expected to be speaker depend and, to some respect, context dependent; the discriminant function used for recognition must be able to take that into account. This led us to compare results obtained from several of them.

In the recognition experiment each of the eight male and eight female speakers prononced nearly eighteen times the same vowel so the total number was 2875. Each of these vowels were either isolated (956), contained in a polysyllabic word (959) or in a phrase (960). For each one two frames of 51.2ms (512 samples) were analysed using an order 12 linear prediction model. Recognition experiments were conducted with five different discriminant functions. As shown by the results recorded in Fig. 1., efficiency of the model appeared for only one of them based on maximum likelihood. Then, even though they are computed from autocorrelation coefficients, predictor coefficients and reflection coefficients yielded much better results. Recognition rates of 82.3% and 80.3% were obtained respectively with predictor coefficients and reflection coefficients but only 62.7% with autocorrelation coefficients. In experiments using the other discriminant functions, autocorrelation coefficients led to recognition rates of about 55% which was slightly prevailing or equivalent to the results obtained with the other types of parameters.

It also appears to be possible to increase the recognition rate by dividing each phoneme class in two, one corresponding to phonemes prononced by male speakers and the other by female speakers. The increase spreads out from 2.6% to 6.2%, it is most important when vowels are represented by reflection coefficients.

LES PARAMETRES

Nous avons choisi une description des voyelles non nasalisées de la langue française définie par trois types de paramètres calculés lors d'une analyse utilisant le modèle linéaire présenté par ITAKURA, F. et SAITO, S. (1968): les coefficients prédicteurs, les coefficients de réflexion et les coefficients d'autocorrélation. Ces derniers contiennent toute l'information retenue relative au signal; les deux premiers types de paramètres y ajoutent ce qui a trait au modèle linéaire.

LES DONNEES

En ce qui concerne notre expérimentation, le corpus utilisé permet l'étude de dix voyelles non nasalisées de la langue française. Il contient 2875 occurences de voyelles, chaque voyelle étant prononcée dix huit fois par chacun des seize locuteurs. Ces voyelles ont été prononcées soit isolément (956), soit à l'intérieur de mots polysyllabiques (959), soit encore à l'intérieur d'une phrase (960). La répartition dans les différents environnements est, en pratique, assurée à parts égales.

Il est admis que l'information relative aux voyelles orales est contenue dans la bande des fréquences comprises entre 50 et 4500 Hz environ. Le traitement du signal a, de ce fait, pour but de conserver cette information, tout en limitant le bruit et la distorsion introduits par les différents traducteurs, et la numérisation, de fournir une suite d'échantillons respectant le théorème de SHANNON. La bande passante utilisée est de fait limitée à l'intervalle compris entre 20 et 4500 Hz.

L'échantillonnage étant réalisé à la fréquence de 10 kHz, ces limitations résultent de ce que nous avons utilisé 512 échantillons du signal pour l'analyse et du filtrage passe-bas, effectué avant la numérisation, qui se caractérise par une fréquence de coupure à 4500 Hz et une atténuation moyenne de 70 dB/oc.

CONDITIONS EXPERIMENTALES

Acquisition

Pour ce qui est des conditions expérimentales, l'onde de pression acoustique fut recueillie en chambre sourde par un microphone Bruël et Kjaer et enregistrée sur un magnétophone Revox A77 dont la bande passante mesurée va de 18 à 16.000 Hz. Le niveau d'enregistrement fut choisi de manière à ne pas dépasser un niveau inférieur de 6 dB à celui pour lequel est observée une distorsion de 1% dans la bande utilisée. Ce dernier est déterminé par essai préalable pour chacune des bandes magnétiques.

Numérisation

L'amplitude du signal enregistré varie d'un locuteur à l'autre et, pour un même locuteur, d'une séquence à l'autre. Le fait est fâcheux pour deux raisons:
- d'abord, il importe de ne pas dépasser la tension maximale admise à l'entrée du convertisseur analogique numérique sous peine d'écrêter le signal;
- ensuite, la conversion entraîne un codage en 9 bits plus un bit de signe qui varie de -512 à +511 pour une excursion en tension de -1 volt à +0,998 volts, et introduit une erreur absolue de 1,95 millivolt. Cela étant, le rapport signal/bruit de numérisation est fonction de l'amplitude du signal.

Ces considérations nous ont conduit à effectuer une égalisation des amplitudes maximales des différentes séquences. Pour ce faire, nous avons dans un premier temps relevé et mémorisé les amplitudes maximales. Dans un second temps et

pour chacune des séquences, nous avons réenregistré le signal en amenant son amplitude maximale à un niveau fixé au moyen d'un atténuateur asservi à l'amplitude maximale mémorisée.

A l'entrée du convertisseur, la valeur maximale de l'amplitude a été fixée à 0,8V ce qui correspond à un rapport signal/bruit de numérisation de 52dB.

ANALYSE NUMERIQUE

En outre, on opère sur le signal numérisé une sélection des segments soumis à l'analyse; elle a pour but d'assurer la correspondance du signal avec le phonème qu'il représente et a pour résultat le positionnement du premier échantillon de ce segment.

Les segments sont alors déterminés par une fenêtre de Hamming. Ils subissent ensuite une préaccentuation qui se présente sous la forme d'un filtrage inverse donné par $P(z) = 1 - \eta z^{-1}$ où η est égal au rapport du second au premier terme de la fonction d'autocorrélation à court terme du signal.

On calcule au moyen de l'algorithme de PFEIFER, L.L. (1973) les 13 premiers termes de la fonction d'autocorrélation discrète et on procède à l'évaluation des coefficients prédicteurs et des coefficients de réflexion (MARKEL, J.D., GRAY, A.H., 1976). La procédure d'analyse fournit, pour chaque segment, 8 vecteurs de représentation affectés d'une étiquette identifiant le locuteur, son sexe et le phonème correspondant. Ces vecteurs sont:

- les coefficients d'autocorrélation normalisés (r_i),
- les coefficients de réflexion (k_i),
- les coefficients prédicteurs (a_i) pour les modèles d'ordre pair de 2 à 12.

L'ordre 12 du modèle étant choisi pour tenir compte du nombre maximum de résonnance pour la bande de fréquences utilisées et d'une correction relative à l'influence des pôles d'ordre supérieur.

SYSTEME DE RECONNAISSANCE

Les systèmes de reconnaissance mis au point utilisent une représentation paramétrique des classes et ne diffèrent que par la fonction discriminante (DIDAY, E., SIMON, J.C., 1976). Chacun utilise une fonction particulière

- basée sur le maximum de vraisemblance

$$D_1(\underline{X},\omega_j) = - \log |K_j| - \left[(\underline{X}-\underline{M}_j)^T K_j^{-1} (\underline{X}-\underline{M}_j) \right]$$

où $\underline{M}_j$ est le vecteur moyen de la classe ω_j dont la matrice de covariance est K_j;

- fondée sur un indice de corrélation

$$D_2(\underline{X},\omega_j) = \frac{\sum_{i=1}^{p} (x_i - cg_i)(m_{ji} - cg_i)}{\left[\sum_{i=1}^{p} (x_i - cg_i)^2 \cdot \sum_{i=1}^{p} (m_{ji} - cg_i)^2 \right]^{1/2}}$$

où les m_{ji} sont les coordonnées de $\underline{M}_j$ et les cg_i celles du centre de masse du nuage dans l'espace de représentation;

- déterminée à partir d'une distance pondérée

$$D_3(\underline{X},\omega_j) = \sum_{i=1}^{p} w_{ji}(x_i - m_{ji})^2$$

où $w_{ji} = 1/\omega_{ji}^2$, ω_{ji}^2 étant l'écart-type de la composante m_{ji} du vecteur moyen;

- définie à partir d'une distance en métrique "city block"

$$D_4(\underline{X},\omega_j) = \sum_{i=1}^{p} |x_i - m_{ji}|$$

- ou tirée d'une distance en métrique euclidienne

$$D_5(\underline{X},\omega_j) = \sum_{i=1}^{p} (x_i - m_{ji})^2$$

Cette dernière fonction est utilisée uniquement en liaison avec l'apprentissage comportant une analyse en composantes principales.

Les systèmes rappelent les résultats de l'apprentissage nécessaire au calcul de la fonction discriminante et les résultats de l'analyse, en un ou deux groupes selon que l'apprentissage distinguait ou non les locuteurs masculins et féminins. Pour chacun des vecteurs $\underline{X}$ de représentation, ils calculent 20 valeurs de la fonction discriminante dans le premier cas, 10 dans le second. Ils recherchent le maximum des fonctions données aux deux premières équations ou le minimum des trois dernières. Ils comparent l'étiquette du vecteur d'entrée avec celle de la classe ainsi retenue et apprécient s'ils ont correctement reconnu ou non le phonème décrit. Ils établissent une matrice de confusion dont les éléments indiquent le nombre de fois qu'un objet affecté d'une étiquette (lignes) a été identifié comme appartenant à une classe de phonèmes déterminée à l'apprentissage (colonnes).

La séparation des locuteurs en deux classes conduit à une structure particulière des matrices de confusion. A la Fig. 2, pour chaque matrice, nous notons en haut et à gauche deux suites de symboles correspondant aux différentes voyelles étudiées. Dans les deux cas, les premières se rapportent aux locuteurs féminins et les secondes aux locuteurs masculins. Cela permet une séparation en quatre parties des matrices de confusion. Les éléments du quartier supérieur gauche donnent le nombre de fois que des voyelles prononcées par des locuteurs féminins ont été reconnues appartenir à des classes définies à partir de locuteurs féminins. Réciproquement ceux du quartier inférieur droit donnent le nombre de fois que des voyelles prononcées par des locuteurs masculins ont été reconnues appartenir à des classes définies à partir de locuteurs masculins. D'autre part, les éléments du quartier supérieur droit donnent le nombre de fois que des voyelles prononcées par des locuteurs féminins ont été reconnues appartenir à des classes définies à partir de locuteurs masculins. Inversément, ceux du quartier inférieur gauche donnent le nombre de fois que des voyelles prononcées par des locuteurs masculins ont été reconnues appartenir à des classes définies à partir de locuteurs féminins.

Lorsque le système a traité tous les objets à reconnaître, il calcule le taux de reconnaissance pour chacune des classes, pour l'ensemble des locuteurs féminins, pour l'ensemble des locuteurs masculins et le taux global.

RESULTATS

Nous avons regroupé sur le tableau de la Fig. 1 l'ensemble des taux de reconnaissance obtenus lors des différentes expériences (RENE de COTRET, Y., 1977). Ils sont présentés selon le type des paramètres et la fonction discriminante utilisée. Sauf lorsque l'on signale une seule classe de locuteurs, l'apprentissage donne pour chaque classe de voyelles deux représentations paramétriques définies d'une part à partir de voyelles prononcées par des locuteurs féminins et d'autre part à partir de voyelles prononcées par des locuteurs masculins.

Ensemble du corpus

Lorsque l'on utilise l'ensemble du corpus, les fichiers sont identiques en apprentissage et en reconnaissance. On remarquera sur le tableau que

- la fonction discriminante basée sur le maximum de vraisemblance permet des taux de reconnaissance nettement plus élevés avec les coefficients prédicteurs (82,3%) et les coefficients de réflexion (80,3%) qu'avec les coefficients d'autocorrélation (62,7%);

- seule cette fonction permet d'obtenir de meilleurs résultats avec les coefficients prédicteurs qu'avec les coefficients d'autocorrélation. On observe un renversement des performances comparées de ces paramètres lorsqu'on passe des méthodes plus simples à celles du maximum de vraisemblance.

On notera également que les quatre dernières fonctions conduisent à des résultats voisins et que la méthode du maximum de vraisemblance ne permet des résultats nettement supérieurs que pour les coefficients issus du modèle linéaire, la différence n'étant que de 3,6% pour les r_i.

En outre, ce tableau indique les gains obtenus par la séparation des classes de locuteurs. Avec cette séparation, on passe de 77,0% à 82,3% pour les a_i, de 74,1 à 80,3% pour les k_i et de 59,1% à 62,7% avec les r_i. Y apparaît le résultat obtenu avec les coefficients de réflexion pour un locuteur et une locutrice; celui-ci révèle une dépendance vis-à-vis du locuteur puisque l'on obtient un taux de 98,2% dans ces conditions.

Voyelles isolées

Le gain obtenu par la séparation des classes de locuteurs est confirmé par les résultats obtenus pour des voyelles prononcées isolément. La première série de résultats concerne des fichiers identiques en apprentissage et en reconnaissance; pour la seconde, les locuteurs sont les mêmes mais les occurences différentes. Dans le premier cas, les résultats portent sur 1912 segments, dans le second sur 956. Pour les a_i, le gain obtenu est de 2,6% ou nul; pour les k_i, ils sont de 5,7% et de 4,7%.

Séparation des locuteurs en deux classes

On pourrait être amené à penser que le gain obtenu n'est imputable qu'au fait de la séparation des locuteurs en deux classes. Cela est vrai en partie, mais les résultats exposés aux Fig. 2 et 3 montrent que le taux de reconnaissance est plus élevé pour des classes définies suivant leur sexe que pour des classes définies arbitrairement. A la Fig. 2 nous présentons deux matrices de confusion obtenues avec les a_i, celle de gauche correspond à un regroupement des locuteurs suivant leur sexe et, pour celle de droite, chaque groupe comprend quatre hommes et quatre femmes; il en est de même à la Fig. 3 pour les k_i. A la Fig. 4, pour les r_i, la première est placée au-dessus de la seconde.

Deux ensembles de résultats sont remarquables. En premier, lorsque les voyelles sont représentées par les coefficients issus du modèle linéaire, les résultats sont plus élevés avec des classes de locuteurs définies suivant leur sexe, respectivement 82,3% contre 79,7% pour les a_i, et 80,3% contre 76% pour les k_i. Il n'en est pas ainsi avec les r_i, on obtient 62,7% contre 62,3%. En second, le rapport de la somme des éléments des quartiers supérieur droit et inférieur gauche d'une matrice de confusion à la somme de tout ses éléments définit le taux d'erreur du système de reconnaissance sur le groupe d'appartenance des locuteurs, c'est-à-dire le taux d'erreurs sur le sexe du locuteur si ces

groupes sont constitués à partir d'éléments masculins d'une part et féminins d'autre part. Il est de 2,9%, 2,8% et de 12,8% respectivement pour les a_i, les k_i et les r_i; mais il est de 23,8%, 21,1% et de 35,1% respectivement si chaque groupe comprend 4 hommes et 4 femmes.

La séparation locuteurs masculins-locuteurs féminins est donc optimale pour la reconnaissance. En outre, les paramètres issus du modèle linéaire conduisent à une représentation des voyelles prononcées par les femmes, sensiblement différente de celle correspondant aux locuteurs masculins. Puisque ces paramètres sont mathématiquement indépendants de la longueur du modèle acoustique du conduit vocal qu'ils représentent (WAKITA, H., 1977), cela peut être rapproché du fait que le rapport des dimensions de la cavité bucale à celle du pharynx est différent chez l'homme et la femme (FANT, G., 1975).

CONCLUSIONS

Ces résultats permettent de constater que l'introduction du modèle linéaire apporte une meilleure description des voyelles, mais que l'on ne peut en tirer profit pour la reconnaissance que si l'on utilise une fonction discriminante rendant compte exactement de la dispersion des paramètres.

Il est en outre possible d'obtenir une amélioration du taux de reconnaissance si l'étape d'apprentissage distingue les locuteurs masculins des locuteurs féminins.

Enfin, dans ce cas, un seul segment de 51,2 ms permet de définir le sexe du locuteur avec une probabilité d'erreur inférieure à 3%.

REFERENCES

DIDAY, E., SIMON, J.C., 1976, Clustering Analysis, in Digital Pattern Recognition, edt Fu, K.S.. - Springer-Verlag, Berlin-Heidelberg-New York, chap. 3, pp. 47-94.

FANT, G. 1975, Nonuniform vowel normalisation, - Speech Transmission Lab., Royal Inst. Tech., Stockholm, Sweden, Quart. Prog. Status Rep., 2-3, pp. 1-19.

ITAKURA, F., SAITO, S., 1968, Analysis Synthesis Telephony Based on the Maximum Likelihood Method. - Procc. of the 6th International Congress on Acoustics, Tokyo, 21-28 Août, C-5-5, pp. C-17-C-20.

MARKEL, J.D., GRAY, A.H. Jr., 1976, Linear Prediction of Speech, Springer-Verlag, Berlin- Heidelberg-New York, 288p.

PFEIFER, L.L., 1973, Multiplication Reduction in Short-Term Autocorrelation. - I.E.E.E. Trans. Audio Electroacoust. Au21, n° 6, pp. 556-558.

RENE de COTRET, Y., 1977, Contribution à la reconnaissance automatique des voyelles orales de la langue française, Thèse de Doctorat, Université de Liège.

WAKITA, H., 1977, Normalisation of Vowels by Vocal-Tract Length and Its Application to Vowels Identification, - I.E.E.E. Trans. Acoustics Speech Signal Processing, ASSP-25, n° 2, pp. 183-192.

Méthode	ordre	Coeff. prédicteurs: a_i				Coeff. de réflexion: k_i					Coeff. d'autocorrélation: r_i			
		Loc. fém.	Loc. masc.	Global	1 classe de loc.	Loc. fém.	Loc. masc.	Global	1 classe de loc.	1 loc. f.+ 1 loc.m.	Loc. fém.	Loc. masc.	Global	1 classe de loc.
Ensemble du corpus														
Maximum de vraisemblance	12	82,9	81,7	82,3	77,0	80,9	79,8	80,3	74,1	98,2	63,6	61,8	62,7	59,1
	10	80,4	79,5	79,9		76,7	76,8	76,8			63,8	67,1	65,4	
	8	76,8	74,9	75,8		72,3	71,8	72,1			61,4	62,1	61,8	
	6	71,1	66,5	68,8		65,9	64,1	65,0			56,6	59,6	58,1	
	4	60,4	52,9	56,6		57,9	52,4	55,1			51,6	48,5	50,1	
	2	41,9	38,2	40,0		39,3	37,5	38,4			36,4	36,6	36,5	
Distance pondérée	12	52,1	44,0	48,1		62,6	60,4	61,5			50,8	57,2	54,2	
Corrélation	12	48,3	41,2	44,7		56,3	58,5	57,5			57,6	60,7	59,1	
Distance en métrique "city block"	12	53,9	43,8	48,9		57,3	56,0	56,6			49,3	53,5	51,4	
Transformation K-L et distance simple	12	46,7	40,1	43,4		49,4	53,6	51,5			47,8	56,6	52,2	
		Taux d'inertie: 87,3				Taux d'inertie: 79,1					Taux d'inertie: 94,0			
Voyelles isolées														
Maximum de vraisemblance														
Fichiers identiques	12	97,4	97,2	97,3	94,7	97,7	94,2	96,0	90,3					
Fichiers différents	12	92,9	89,7	91,3	91,6	90,8	84,4	87,6	82,9					

Fig. 1. ENSEMBLE DES RESULTATS

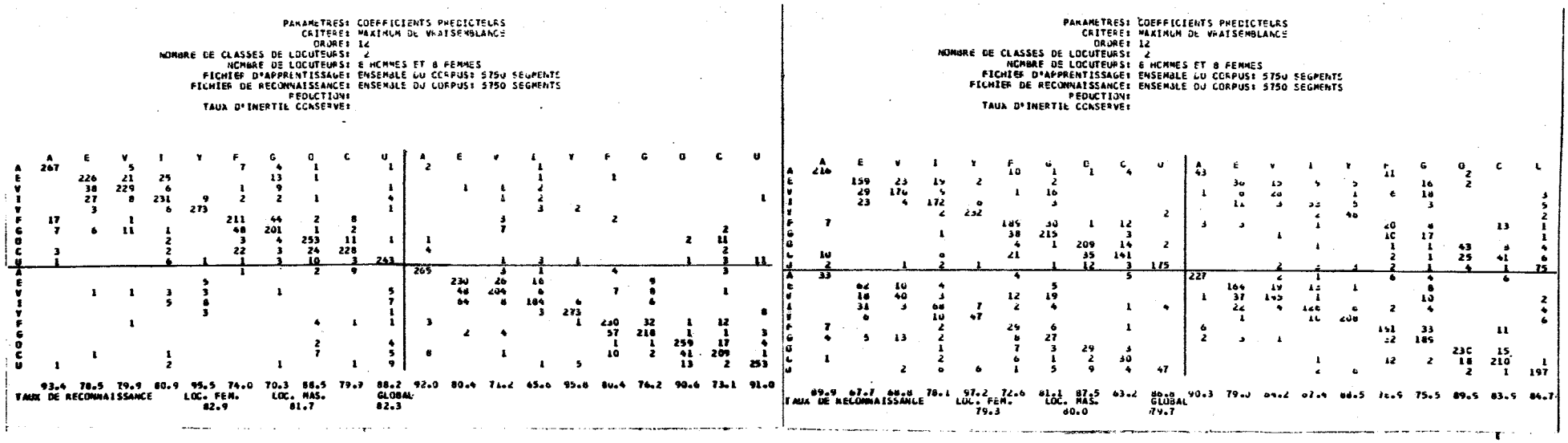


Fig. 2. COEFFICIENTS PREDICTEURS

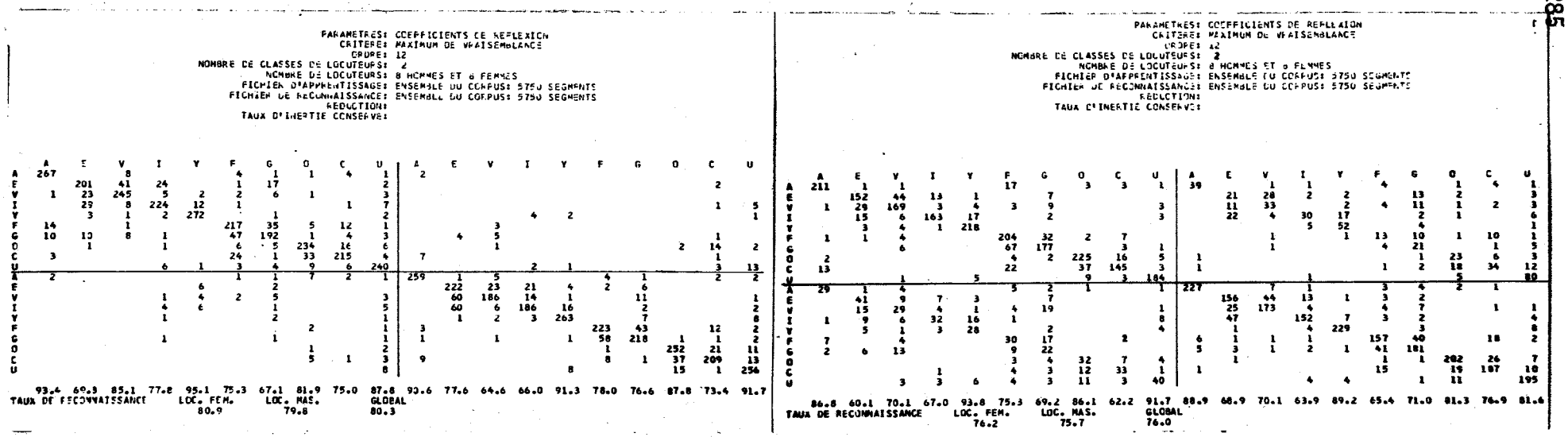


Fig. 3. COEFFICIENTS DE REFLEXION

PARAMETRES: COEFFICIENTS D'AUTOCORRELATION
CRITERE: MAXIMUM DE VRAISEMBLANCE
ORDRE: 12
NOMBRE DE CLASSES DE LOCUTEURS: 2
NOMBRE DE LOCUTEURS: 8 HOMMES ET 8 FEMMES
FICHIER D'APPRENTISSAGE: ENSEMBLE DU CORPUS: 5750 SEGMENTS
FICHIER DE RECONNAISSANCE: ENSEMBLE DU CORPUS: 5750 SEGMENTS
REDUCTION:
TAUX D'INERTIE CONSERVE:

	A	E	V	I	Y	F	G	O	C	U	A	E	V	I	Y	F	G	O	C	U
A	246		7			8	3	1	5	2	9			1				5		1
E		158	85	10	2		8	1	1					1				10	12	
V	4	21	420	2	1		5			1				1				1	23	3
I		12	10	125	62	2	1			30			1	4	1		1	30	4	5
Y		3	1	9	161		1			7			1	12	11			73		9
F	42		5	1		121	19	6	60	1	14		3			6	1	6	3	
G	12	4	61	1		61	72	6	35			3	16	3			1	8	2	1
O		1				3	3	221	13	2	1						2	20	20	2
C	5		1			10	5	40	168		12							19	28	
U			1	5	2	2		5	3	124				2	1		2	11	1	129
A	10					2	1	1	15		238		1			6	2	5	6	1
E			1		12	2	1					193	30	18		1	24		2	2
V	1	2	4	1	3	7	1		5	2	8	32	179	6	1	10	22		1	3
I	1		4	4	26		1			6		51	27	125	18		8	14		3
Y				2	4							1		14	223		5	14		25
F								2			6		1			143	31	72	29	2
G	1			1						1	1	7	3	1		58	165	12	7	29
O			2					11		4	1					2		30	190	48
C	2					1	1	18	9	5	10					7	2	23	203	5
U	1			3			1			16				2	5	1	2	39		218
	88.5	54.9	78.5	44.8	59.7	44.1	25.5	83.7	68.1	87.8	86.1	67.5	63.5	44.8	78.8	50.0	57.7	14.2	74.1	81.3

TAUX DE RECONNAISSANCE — LOC. FEM. 63.6 — LOC. MAS. 61.8 — GLOBAL 62.7

PARAMETRES: COEFFICIENTS D'AUTOCORRELATION
CRITERE: MAXIMUM DE VRAISEMBLANCE
ORDRE: 12
NOMBRE DE CLASSES DE LOCUTEURS: 2
NOMBRE DE LOCUTEURS: DEUX GROUPES COMPRENANT CHACUN 4 HOMMES ET 4 FEMMES
FICHIER D'APPRENTISSAGE: ENSEMBLE DU CORPUS: 5750 SEGMENTS
FICHIER DE RECONNAISSANCE: ENSEMBLE DU CORPUS: 5750 SEGMENTS
REDUCTION:
TAUX D'INERTIE CONSERVE:

	A	E	V	I	Y	F	G	O	C	U	A	E	V	I	Y	F	G	O	C	U
A	191		3			10		4	20	2	36					11			6	5
E	2	100	72	20	9	12	1	5	6		2	17	18	5	6	1	4	5	2	1
V	4	24	148	3	2	38	5	2	6	2	3	2	30	4	3	1	6	3	3	2
I	1	13	11	113	56		3	4	1	13		8	4	27	17		1	2		14
Y	1	1	4	7	216					14		2		7	26					10
F	16		2			134	9	16	48	1	3					12	7	7	33	
G	3	1	13	2		68	64	6	37	13	4	8	5	7	1	6	20	3	14	10
O						2	2	241	14	10				1		1		13	4	
C	12		1			17	2	87	112	1	3					3		5	39	6
U			3		1		4	6	1	253				2	6			2		13
A	126		3			1	2		10	2	129		1			5	5		3	
E		22	30	10	5	7	5	2	4			97	60	23	6		3	4	3	
V	1	12	37	5	1	29	14	9	7	2	6	12	125	1	4		5	8	8	2
I	1	3	24	29	54		4	3	2	18		15	2	85	20	2	5			11
Y			1	16	114		1			11		2		11	114		3			13
F	44		5	3		100	7	1	14	1	23		1	1		41	25	1	19	
G	12	3	45	3		68	4		4	1	10	3	10	4	3	17	92		7	
O				1		5	3	154	20	3	1					1		91	8	1
C	14		1			14	5	47	54		1					5		31	114	
U	1		3	1	3	3	4	5	6	184				1	3		3	13		58
	78.0	40.6	61.8	48.6	84.0	50.7	29.4	88.2	52.4	92.4	88.5	41.6	56.3	43.1	79.2	49.3	33.6	85.1	58.7	84.0

TAUX DE RECONNAISSANCE — LOC. FEM. 62.7 — LOC. MAS. 62.0 — GLOBAL 62.3

Fig. 4. COEFFICIENTS D'AUTOCORRELATION

9èmes JOURNEES D'ETUDE SUR LA PAROLE

LANNION 31 mai - 2 juin 1978

LA DESCRIPTION AU NIVEAU ACOUSTIQUE DES CONSONNES NASALES PRONONCEES DANS UN DISCOURS CONTINU

Ryszard GUBRYNOWICZ - Institut des Recherches Fondamentales de Technologie Academie Polonaise des Sciences - Varsovie - Pologne

Renato DE MORI - Istituto di Scienze dell'Informazione - Università di Torino - Torino - Italie

Pietro LAFACE - CENS - IENGF. Istituto di Elettrotecnica - Politecnico di Torino - Torino - Italie

.RESUME

Un système de description acoustique des consonnes nasales est prevu comme source d'informations pour un systeme de comprehension de la parole continue. Les recherches étaient limitées aux deux nasales seulement, bilabiale /m/ et alveolaire /n/ prononcees dans tous les contextes intervocaliques possibles. La consonne vélaire /ŋ/ n'intervenant pas dans ce contexte n'a pas été l'objet de notre analyse.

Diverses méthodes de caractérisation des consonnes nasales furent étudiées et finalement on a retenu pour leur description trois paramètres:

- deux points caractéristiques de la transition du deuxième formant, la frequence initiale et terminale, déterminées au début et à la fin du segment nasal;
- différence d'énergie entre le premier et deuxième formant déterminee dans le moment où l'energie du deuxieme formant atteint le minimum absolu.

Ces trois paramètres ont servi à mettre au point les règles d'identification de consonnes nasales tenant compte du type de coarticulation caracterisé par le lieu d'articulation de voyelles adjacentes. On a obtenu ainsi un ensemble de 9 règles d'identification basées sur le principe des relations floues. Ces règles furent vérifiées sur un corpus de 200 échantillons provenant des énonces de 4 locuteur et on a obtenu un pourcentage d'erreurs egal à 6%.

ACOUSTIC DESCRIPTION OF NASALS EXTRACTED FROM CONTINUOUS SPEECH

SUMMARY

A methodology of acoustic-phonetic description of nasal consonants uttered in intervocalic contexts in continuous speech is presented. A motivation that the rules for nasal classification should depend on both vowels of the VNV syllable is given. An investigation of their acoustic patterns in VCV utterances extracted from continuous speech may allow to infer rules for describing a variety of coarticulation effects exhibited in different contexts by many speakers.

The investigation whose results are presented in this paper, was limited to the characterization of the bilabial /m/ and the alveolar /n/. The velar /ŋ/ was not considered because it does not occur in such contexts.

The rules inferred after experiments reflect the theoretical basis which assume that the generation of VCV sequences involves the encoding of three discrete phonemes in such a way that the vocal tract shape may change continually during the entire utterance of the syllable exhibiting various degree of the influence of the vowels on the consonant.

Our research confirm that vowel transitions predominated over nasal resonances as a cue for identification of nasal consonants.

The elaborated identification rules take into account the global variations of analysed parameters from the beginning to the end of the analysed syllable. The set of acoustic parameters used for description of nasals was composed of: initial frequency F_{21} and terminal frequency F_{22} of the second formant transition during the articulation of the consonant part of the syllable and difference of formants energy D_{12} /see fig. 1/.

The vocalic contexts are represented by places of articulation determined on the basis of the values assumed by the second formant on the stationary portion of the vocalic segment according to the formula /1/.

A concise representation of the nasals based on the usage of fuzzy relations is obtained by the answers to the questionnaire whose composition is given in the table 1. Each particular question Q_{Ni} /component question/ concern the variations of a given acoustic parameter. The answer to the component question is fuzzy linguistic function of the variables h_{Ni}, l_{Ni} defined by the table 1.

Formulas /2/ and /3/ give the rules for generating hypotheses about intervocalic nasal consonants. Such rules account coarticulation effects and allow us to classify a set of 200 samples from 4 male speakers with the overall error rate of 6%.

LA DESCRIPTION AU NIVEAU ACOUSTIQUE DES CONSONNES NASALES PRONONCEES DANS UN DISCOURS CONTINU

Ryszard GUBRYNOWICZ - Institut des Recherches Fondamentales de Technologie Academie Polonaise des Sciences - Varsovie - Pologne

Renato DE MORI - Istituto di Scienze dell'Informazione - Università di Torino - Torino - Italie

Pietro LAFACE - CENS - IENGF. Istituto di Elettrotecnica - Politecnico di Torino - Torino - Italie

1. INTRODUCTION

L'importance de la reconnaissance efficace des consonnes nasales pour l'amélioration des performances du systeme de comprehension de la parole ne laisse aucun doute, vu leur frequence d'occurrence qui atteint pour beaucoup de langues env. 11% /p.ex. pour l'anglais, italien, polonais et autres/.

Comme dans le cas de sons vocaliques le spectre des consonnes nasales a aussi une structure formantique mais determinee par les proprietes acoustiques de trois cavites conjuguees: orale, nasale et pharyngale. Les resultats obtenus par plusieurs chercheurs /MALECOT, 1956; NAKATA, 1959 ; FUJIMURA, 1962; KACPROWSKI et al., 1965; DUKIEWICZ, 1967; GILLMAN, 1974/ montrent que les formants des nasales sont plus ou moins regulierement repartis dans le spectre et espaces entre eux de 800-1000 Hz. Pour un locuteur donne, leurs frequences à l'état quasistationnaire sont pour toutes consonnes nasales similaires, neanmoins dependent du contexte phonetique.

Dans ce rapport nous décrivons un système de description au niveau acoustico-phonetique des consonnes nasales prevu comme source d'informations pour un systeme de comprehension de la parole. Ce procede concerne les consonnes nasales en position intervocalique et demande une analyse des évolutions des parametres formantiques qui embrasse la prononciation de la syllabe VNV entiere. Une étude prealable des caracteristiques acoustico-phonetiques des consonnes nasales donne un appui theorique à la methode de leur description élaboree. Les resultats de reconnaissance rapportes en conclusions confirment la validite de cette méthode à condition d'utiliser pour l'identification des nasales les relations floues.

2. LES CARACTERISTIQUES ACOUSTIQUES DES CONSONNES NASALES

De nombreux travaux ont été entrepris dans le but d'établir la meilleure description acoustique possible des consonnes nasales /p.ex. en outre des auteurs cités dans l introduction - HECKER, 1962; GARCIA, 1967 / mais les résultats y apportés sont souvent incompatibles.

Néanmoins, la plus part de chercheurs est d'accord que la fréquence du premier antiformant /N1/ est le trait essentiel qui permet distinguer les nasales entre eux car elle dépend principalement de l emplacement du point où a lieu la fermeture complete du tube buccal /FUJIMURA, 1962; HECKER, 1962/. La frequence N1 baisse quand le lieu d'articulation se déplace vers l'avant de la cavite buccale. Il est généralement admis que la frequence de l antiformant de la nasale bilabiale /m/ est env. 750-1250 Hz

de la nasale alvéolaire /n/ - 1450÷2200 Hz et de la vélaire /ŋ/ dépasse 3000 Hz.

En dépit de variations importantes de la fréquence de l'antiformant pour les consonnes nasales, il est en effet difficile d'utiliser ce parametre pour les caractériser. Il arrive souvent que celui-ci n'est pas facile à detecter sans équivoques dans le spectre de la parole naturelle. Cela est dû a la complexite de l'effet du shunt de la cavite buccale.

L'influence de cet effet sur le spectre final de la consonne nasale n'est pas toujours évidente a cause du dephasage de l'impedance d'entree de la cavite buccale qui est difficile a évaluer. En plus, il arrive que l'antiformant dont le facteur Q est relativement bas se situe à proximite du deuxieme ou troisième formant ce qui entraîne leur amortissement ou même la disparition de l'un ou de l'autre.

Dans cette situation même les méthodes le plus raffinées qui demandent un nombre de calculs tres important, comme prédiction lineaire inverse et prédiction homomorphique /KOPEC et al., 1977/ ou celle recemment proposée par STEIGLITZ /1977/ donnent des valeurs des zeros qui souvent ne correspondent pas aux valeurs reelles des antiformants, même dans les cas où ils sont faciles a déceler dans le spectre de la nasale.

Certains résultats de recherches avec la parole synthétique montrent que de point de vue perceptionnel la valeur absolue de la fréquence de N1 n est pas l'indice majeur /NAKATA, 1959; KACPROWSKI et al., 1965/. Ce surtout la position de l'ensemble /cluster/ F2-N1-F3 qui joue un rôle essentiel. On a obtenu une bonne synthese des nasales en le remplacant par un formant d'une largeur de bande relativement grande. Les frequences typiques de ce formant equivalent sont: pour /m/ - 1100 Hz et pour /n/ - 1700 Hz.

Il est intéressant à noter que l'analyse expérimentale du signal naturel donne au deuxième formant des valeurs similaires a ceux que nous venons de citer. Mais cette frequences est susceptible aux variations aléatoires relativement importantes et en general la distribution des formants ne permet pas la distinction suffisante des nasales entre elles /voir p.ex. les résultats obtenu par GILMAN, 1974/.

Bon nombre de travaux indiquent que les informations sur les consonnes nasales ne sont pas seulement codées dans la distribution des formants mais aussi, et probablement avant tout, dans les évolutions des phases transitoires entre le segment nasal et les sons adjacents. L'importance des phases transistoires est particulierement évidente dans le cas du contexte vocalique. Il est bien connu, par exemple, que les auditeurs commettent beaucoup d'erreurs dans l'identification des nasales extraites de son entourage phonetique et dépourvues de phases transitoires /MALECOT, 1956; DUKIEWICZ, 1967/. Il faut souligner qu'une experience similaire avec des consonnes fricatives voisees ne pose aucun probleme pour les auditeurs.

Il parait que le paramètre le plus efficace pour la classification des est la frequence initiale de la transition du deuxieme formant. Les expériences avec les syllabes synthétisees N-V /NAKATA, 1959/ prouvent que cette frequence est relativement basse /entre 800 a 1300 Hz/ quand la nasale est identifiée comme /m/ et elle prend une valeur plus haute /1500 ÷ 2000 Hz/ quand il s'agit de /n/. En plus, le même auteur a démontré que les limites superieures concernent plutôt les cas où la voyelle adjacente est /i/ tandis que les limites inférieures vont mieux aux syllabes avec la voyelle /u/.

Il est à noter qu'on a comparé l'importance des transitions, celle qui précède le segment consonantique avec celle qui le suit, pour l'identification de cette consonne /BRADY et al., 1961/. On a constaté que la transition qui précède ce segment le caractérise dans une proportion plus forte que la transition qui le suit. Dans le cas particulier des consonnes nasales une explication de ce fait vient d'une analyse du processus d'articulation. En effet, MOLL /1962/ a démontré que le moment de l'ouverture de la porte vélique anticipe bien plus le début de l'articulation du segment nasal, tandis qu'elle se ferme avec un petit retard seulement, après la fin de ce segment.

En plus, il faut souligner que grâce à l'effet de coarticulation la transition dans la partie initiale de la syllabe VNV dépend aussi de la seconde voyelle. On observe que les mouvements des formants vers les valeurs terminales commencent presque simultanément avec le début de l'articulation de la partie consonantique, ou même un peu plus tôt.

Ce phénomène peut être expliqué en tenant compte du fait que pendant la prononciation de la nasale la partie de la cavité buccale qui est en arrière du point d'articulation peut adopter la configuration correspondante à l'articulation d'une des voyelles adjacentes.

Alors, pour la mise au point d'une description des nasales efficace il faut considérer les variations globales des paramètres au cours de la prononciation de la syllabe entière.

3. LE CORPUS ANALYSE

Le corpus enregistré par 4 locuteurs masculins était constitué d'une cinquantaine de phrases qui ont été formées de manière à obtenir toutes les coarticulations possibles des nasales /m/ et /n/ avec les cinq voyelles /i/, /e/, /a/, /o/ et /u/. En somme, pour chaque locuteur on a analysé env. 50 échantillons. Une attention spéciale était donnée aux nasales prononcées dans l'entourage de voyelles antérieures car ces cas ont apparu comme le plus difficiles du point de vue de l'analyse à l'aide de la technique de poursuite des formants.

4. DETECTION DES INDICES ACOUSTICO-PHONETIQUES DES NASALES

L'acquisition du signal enregistré en ambiance peu bruyante était effectuée à l'aide d'un convertisseur A/D de 10 bits avec fréquence d'échantillonage 20 kHz. Après on a déterminé les variations du pitch et on a calculé les spectres FFT et LPC synchrones. L'étape suivante du traitement donnait les variations de l'énergie totale et de l'énergie dans les bandes de fréquences 200-900 Hz, 3-5 kHz et 5-10 kHz. Le traitement de données se terminait par une procédure de poursuite des formants.

Ce jeu de paramètres a servi à élaborer l'algorithm d'extraction de pseudo-syllabes VNV /DE MORI et al., 1977a/.

Pour la description des nasales nous avons utilisé les paramètres suivants:

- $F_i(nT)$: fréquence du ième formant déterminée pour l'échantillon n / T - période d'échantillonage/;
- $A_i(nT)$: niveau d'énergie du ième formant de l'échantillon n;

- $D_{12} = A_1(n^*T) - \text{Max}\{A_2(n^*T), A_3(n^*T)\}$ où n^* correspond au échantillon pour lequel $A_2(nT)$ atteint le minimum absolu entre la première et seconde voyelle;

- F_{21} : fréquence initiale du deuxième formant déterminée au début du segment nasal;

- F_{22} : fréquence terminale du deuxième formant à la fin du segment nasal.

La fig. 1 montre les évolutions des formants /1a/ et de leurs niveaux /1b/ au cours de la prononciation de la pseudo-syllabe /ine/. On y a marqué les points caractéristiques de ces variations F_{21}, F_{22} et indiqué la difference des niveaux D_{12}.

Cet ensemble des paramètres dépend fortement du contexte voçalique. Le type de la voyelle adjacente était caractérise par son lieu d'articulation que nous avons défini comme antérieur, central ou postérieur selon la frequence moyenne du deuxième formant determinee pour la partie quasi-stationnaire du segment vocalique. Pour la définition du type de la voyelle nous avons choisi les frequences limites du deuxieme formant suivantes:

$$\begin{aligned} 1900\ \text{Hz} < \bar{F}_2 \quad &- \text{pour voyelles antérieures} \\ 1300\ \text{Hz} < F_2 < 1900\ \text{Hz} \quad &- \text{" " centrales} \\ F_2 < 1300\ \text{Hz} \quad &- \text{" " postérieures .} \end{aligned} \qquad (1)$$

Nous ne donnerons ici qu'une très brève esquisse de la procédure de reconnaissance basée sur les relations floues /pour les détails voir DE MORI et al., 1978/. Le processus d'identification consiste à trouver la réponse globale à un questionnaire dont chacune des questions concerne les variations d'un parametre donne. Selon la valeur de ce parametre on évalue le degre de l'appartenance plausible de la nasale analysée à l'une de deux classes. Cette estimation est faite d'après la courbe /membership function/ établie au cours du processus d'apprentissage de façon à éviter les réponses erronees avec un degré de plausibilite égal à un et en minimalisant le nombre de ces réponses dont ce degré dépasse celui des réponses correctes. La prise de la décision définitive sur l'appartenance de la nasale en question est en accord avec la réponse qui a obtenu le degré de plausibilite égal à un ou suffisament haut pour rejeter l'autre alternative.

Le tableau 1 donne les valeurs limites des paramètres utilisés pour la description des nasales. Hors de ces limites le degre de l'appartenance plausible à l'une de deux classes est égal à un. Alors, la valeur H_{Ni} correspond à la valeur du parametre donne qui permet d'accepter l'hypothèse avec certitude /égale à un/ que la consonne analysee est /n/ et de rejeter /avec un degré nul/ l'hypothèse que c'est la consonne /m/. L_{Ni} correspond à une situation inverse. Ni désigne le numero de la question.

Les règles de reconnaissance des consonnes nasales prennent une forme globale des expressions logiques suivantes où les symboles [f], [c] et [b] désignent les voyelles antérieures, centrales et postérieures:

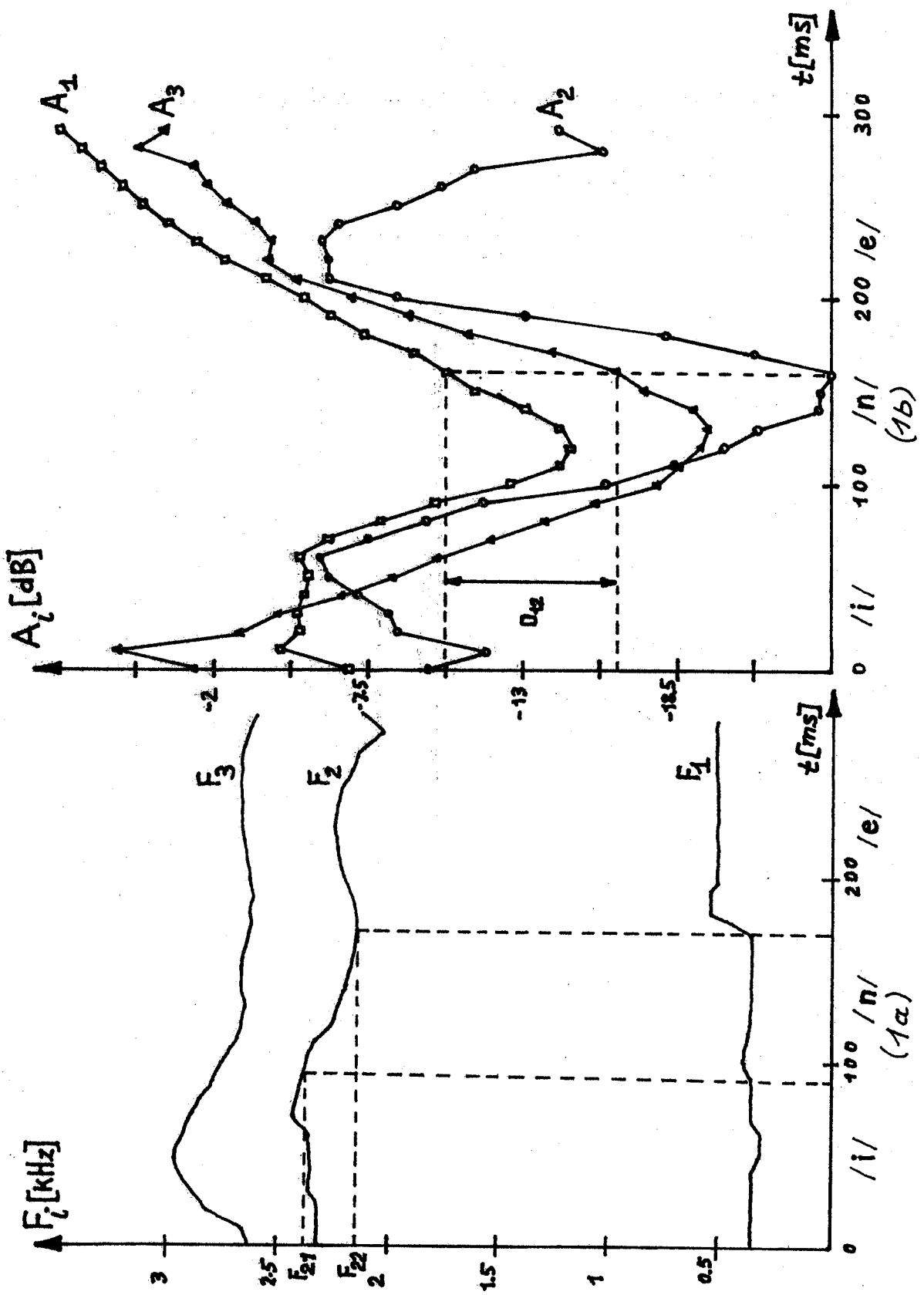

A1
A3
A2
Ai [dB]
-2
-7.5
-13
-18.5
D12
t[ms]
0
/i/
100
/n/
(1b)
200
/e/
300
F3
F2
F1
Fi [kHz]
3
2.5
F21
F22
2
1.5
1
0.5
t[ms]
0
/i/
100
/n/
(1a)
200
/e/

Fig. 1

Tableau 1

Description des questions concernantes les variations des paramètres acoustiques utilisés dans les règles de reconnaissance

Question	p_{Ni}	L_{Ni}	H_{Ni}	Dimension
Q_{N1}	F_{21}	1880	2320	Hz
Q_{N2}	D_{12}	0	15	dB
Q_{N3}	F_{21}	1750	2150	Hz
Q_{N4}	F_{22}	1500	1700	Hz
Q_{N5}	D_{12}	6	17	dB
Q_{N6}	F_{21}	1400	2250	Hz
Q_{N7}	F_{22}	900	1400	Hz
Q_{N8}	D_{12}	9	23.5	dB
Q_{N9}	F_{21}	1500	1700	Hz
Q_{N10}	D_{12}	8	17	dB
Q_{N11}	F_{21}	1300	1500	Hz
Q_{N12}	F_{22}	1300	1500	Hz
Q_{N13}	F_{21}	1300	1480	Hz
Q_{N14}	F_{22}	960	1100	Hz
Q_{N15}	F_{21}	940	1200	Hz
Q_{N16}	F_{22}	900	1100	Hz
Q_{N17}	D_{12}	6	20.5	dB
Q_{N18}	F_{21}	900	1000	Hz
Q_{N19}	F_{22}	980	1400	Hz
Q_{N20}	D_{12}	10	19	dB
Q_{N21}	F_{21}	980	1050	Hz
Q_{N22}	F_{22}	900	1000	Hz

$$\begin{aligned}<m> = &[ff](l_{N1} + l_{N2}) + [fc](l_{N3} + l_{N4} + l_{N5}) + [fb](l_{N6} + l_{N7} + l_{N8}) \\ &+ [cf](l_{N9} + l_{N10}) + [cc](l_{N11} + l_{N12}) + [cb](l_{N13} + l_{N14}) \\ &+ [bf](l_{N15} + l_{N16} + l_{N17}) + [bc](l_{N18} + l_{N19} + l_{N20}) + [bb](l_{N21} + l_{N22})\end{aligned} \quad (2)$$

$$\begin{aligned}<n> = &[ff](h_{N1} + h_{N2}) + [fc](h_{N3} + h_{N4} + h_{N5}) + [fb](h_{N6} + h_{N7} + h_{N8}) \\ &+ [cf](h_{N9} + h_{N10}) + [cc](h_{N11} + h_{N12}) + [cb](h_{N13} + h_{N14}) \\ &+ [bf](h_{N15} + h_{N16} + h_{N17}) + [bc](h_{N18} + h_{N19} + h_{N20}) + [bb](h_{N21} + h_{N22})\end{aligned} \quad (3)$$

Le type de coarticulation pour la règle correspondante est indiqué entre parentheses carrees.

Remarquons enfin /tableau 1/, que les fréquences F_{21} et F_{22}, diminuent quand le lieu d'articulation de voyelles adjacentes se déplace en arriere. Il est interessant de noter que cette tendance existe aussi dans le cas ou il y a un changement d'une de voyelles seulement. Par exemple, la frequence F_{21} qui depend plus fortement de la voyelle qui précède le segment nasal baisse quand la voyelle qui le suit passe de l'antérieure à la posterieure. Ce phenomene illustre bien le fait que la transition entre la voyelle initiale et la consonne depend dans un certain degré aussi de la voyelle finale.

5. CONCLUSIONS

Les résultats de reconnaissance de consonnes nasales obtenus sont appréciables, En effet, le pourcentage d'erreurs moyen était environ 6% et n'a pas dépassé 16% dans le cas de coarticulation voyelle antérieure - consonne nasale - voyelle postérieure pour lequel les changements du deuxieme formant sont les plus rapides.

Il faut souligner que les bons résultats de reconnaissance étaient obtenus grâce à l'introduction aux regles de reconnaissance des informations sur l'entourage vocalique de la nasale, qui se caractérise d'ailleurs par une forme tres simple. A l'état actuel il ne semble pas qu'il soit nécessaire de décrire le contexte avec plus de précision, par exemple, en identifiant les voyelles.

Tous nos remerciements à Mr. W. MIKIEL de l'Institut des Recherches Fondamentales de Technologie à Varsovie pour les remarques sur les proprietés acoustiques de consonnes nasales.

La partie expérimentale de cette recherche était entièrement effectuée à Centro per l'Elaborazione Numerale dei Segnali - CNR à Turin.

REFERENCES

BRADY, P.T., HOUSE, A.S., STEVENS, K.N., 1961, Perception of sounds characterized by a rapidly changing resonant frequency. - J.A.S.A. 33, pp. 1357-1362.

DE MORI, R., LAFACE, P., PICCOLO, E., 1976, Automatic detection and description of syllabic features in continuous speech. - IEEE

Trans. on ASSP 24, n° 5, pp. 365-379.

DE MORI, R., LAFACE, P., TORASSO, P., 1977 a, Automatic recognition of liquids and nasals in continuous speech. - Conf. Rec. of IEEE Int. Conference on Acoustics, Speech and Signal Processing, Hartford, Conn., pp. 644-647.

DE MORI, R., LAFACE, P., SARDELLA, M., 1977 b, Ambiguities, fuzziness and precategorical classification in a speech understanding system. - IEEE Trans. on Computers,

DE MORI, R., GUBRYNOWICZ, R., LAFACE, P., 1978, Inference of a knowledge source for the recognition of nasals in continuous speech. /à paraître/.

DUKIEWICZ L., 1967, Les sons nasals polonais /en polonais/, PWN, Warszawa.

FUJIMURA, O., 1962, Analysis of nasal consonants. - J.A.S.A. 34, pp. 1865-1875.

GARCIA, E., 1967, Identification and discrimination of synthetic nasals. - Haskins Lab. Status Rep. on Speech Res., 10. 8.1 .

GILLMAN, R.A., 1974, Automatic recognition of nasal phonemes. - Proc. IEEE Symp. on Speech Recognition, Pittsburg, pp. 74-79.

HECKER, M.L., 1962, Studies of nasal consonants with an articulatory speech synthesizer. - J.A.S.A. 34, pp. 179-188.

KACPROWSKI, J., MIKIEL, W., 1965/66, Simplified rules for parametric synthesis of nasal and stop consonants in C-V syllables by means of the "terminal-analog" speech synthesizer. - Acustica 16 , n° 6, pp. 356-364.

KOPEC, G.E., OPPENHEIM, A.V., TRIBOLET, J.M., 1977, Speech analysis by homomorphonic prediction. - IEEE Trans. on ASSP 25, n° 1, pp. 40-49.

MALECOT, A., 1956, Acoustic cues for nasal consonants. - Language 32, n° 2, pp. 274-284.

MOLL, K.L., 1962, Velopharyngeal closure on vowels. - J. Speech & Hearing Res. 5, pp. 30-37.

NAKATA, K., 1959, Synthesis and perception of nasals consonants. - J.A.S.A. 31, pp. 661-666.

STEIGLITZ, K., 1977, On the simultaneous estimation of poles and zeros in speech analysis. - IEEE Trans. on ASSP 25, n° 3, pp. 229-234.

9èmes JOURNEES D'ETUDE SUR LA PAROLE

LANNION 31 mai - 2 juin 1978

ETUDE DU TRAIT PERMETTANT DE DISTINGUER ENTRE LES 3 CLASSES DE CONSONNES EXPLOSIVES PB, TD, KG.

Pierre ALINAT THOMSON-C.S.F. Division des Activités Sous-Marines
B.P. 53 - 06802 CAGNES-sur-MER

.RESUME

Il s'agit de déterminer un ensemble de règles permettant de séparer les 3 classes de consonnes PB, TD et KG (en laissant de côté le trait sourde-sonore : PTK/BDG). Cette étude a été menée par analyse de 1836 syllabes CV (34 locuteurs x 6 consonnes x 9 voyelles). Le système d'analyse est une cochlée artificielle. Une étude bibliographique du sujet montre qu'il faut distinguer les évènements suivants : phase soutenue, impulsion et transition vocalique et que le trait étudié ici est surtout localisé au niveau de l'impulsion. Un programme a donc été écrit pour localiser l'impulsion et en extraire un certain nombre de paramètres qui semblaient à première vue pertinents. A partir des résultats obtenus, des règles de séparations simples et indépendantes du trait sourde-sonore ont été définies : elles font intervenir la forme du spectre et la position du F2 de la voyelle associée.

Les matrices de confusion calculées au moyen d'un programme utilisant ces règles, montrent que le taux d'erreur obtenu est de l'ordre de 10 %.

SOME ACOUSTIC FEATURES USED TO DISTINGUISH THE STOP CONSONANTS /PB, TD, KG/.

Pierre ALINAT THOMSON-C.S.F. ASM Division, 06802 CAGNES-sur-MER - France

. SUMMARY

In this paper cues for distinguishing the 3 classes of stop consonants /PB/, /TD/ and /KG/ are investigated (the voice-voiceless feature is not taken into account here). A bibliographic study on this subject shows that stop consonant identification seems to involve 3 acoustic segments : silence, burst, vocalic transition, and that burst carries the largest perceptual load for the /PB, TD, KG/ distinction. In this study 1836 isolated CV syllables (34 speakers x 6 consonants x 9 vowels) have been analyzed by means of an analog cochlea (a bank of 96 band pass filters whose transfer functions and center frequencies are deducted from the human ear). A program has been written to localize bursts and estimate a few pertinent parameters. From the results obtained, rules to distinguish /PB, TD, KG/ have been implemented. These rules are simple and independent from the voice-voiceless feature :

/KG/ is characterized by a narrow spectral peak near the F2 of the adjacent vowel, /TD/ by a narrow spectral peak far above F2 or by a flat spectrum in which the higher frequencies predominate ; /PB/ by a narrow spectral peak under F2 or by a flat spectrum in which the lower frequencies predominate.

A recognition program using these rules has been written and confusion matrix obtained for the 1836 syllables shows that error rate is about 10 %.

La reconnaissance analytique des consonnes explosives s'exécute en trois étapes :

- Localisation
- Classification sourdes (PTK) ou sonores (BDG) ou nasales (M, N)
- Classification labiale (PBM) ou dentale (TDN) ou palatale (K, G)

C'est de cette troisième étape dont il est question ici. Plus précisément, il s'agit d'expliciter quantitativement des règles de classification et de tester leur valeur pour un nombre élevé de locuteurs. Il s'agit d'une étude par analyse. Le capteur est une batterie de 96 filtres passe-bande (+ détection intégration) dont les paramètres sont déduits de ceux de l'oreille humaine. Lorsque l'étude a été menée, la sélectivité des filtres était insuffisante pour permettre de séparer M et N : ces 2 phonèmes ne sont donc pas pris en compte ici.

1. - ETUDES ANTERIEURES.

Parmi les études antérieurement menées sur ce sujet, celles qui ont été utilisées ici sont :

a) Haskins Laboratories:
Il s'agit essentiellement des expérimentations par synthèse bien connues. La notion de locus sur laquelle étaient basées la plupart de leurs manipulations est abandonnée de nos jours. Il faut toutefois noter une synthèse de consonnes explosives au moyen d'impulsions de bruit à bande étroite (Liberman A.M. 1952) dont les résultats seront évoqués au § 4.

b) Halle M. & Al (1957) firent clairement la distinction entre transition vocalique et burst. Il faut entendre par burst l'impulsion de bruit plus ou moins haute fréquence se produisant pendant et juste après l'ouverture du conduit vocal. La transition vocalique consiste en une phase où les 2 formants de la voyelle associée sont présents et évoluent vers les positions propres à la voyelle. Lorsque la consonne explosive est associée à une consonne fricative (par exemple /TCH/ /PR/ etc...) il devient difficile de parler de transition vocalique. Lorsque la consonne explosive est associée à une voyelle, Halle nota l'existence des 2 variétés allophoniques :

phase soutenue - burst - transition vocalique - voyelle (CV)
voyelle - transition vocalique - phase soutenue - burst (VC)

La recherche, au niveau de la transition et du burst, de règles de classification fut faite par analyse de sonagrammes et a fourni des résultats qualitatifs.

c) Winitz H. & Al (1972) et La Rivière C (1975) firent des études portant surtout sur la perception de syllabes isolées (contenant une explosive) plus ou moins tronquées de façon à tester les importances relatives du burst et de la transition vocalique. Leur conclusion est finalement d'une part que la transition vocalique, bien que parfois utile, n'est pas toujours nécessaire pour la classification P ou T ou K, d'autre part que le burst, lui, est absolument nécessaire.

d) Alinat (1975) rechercha qualitativement des traits permettant de séparer en 3 classes PB - TD - KG. Le système d'observation était une cochlée artificielle (96 filtres passe-bande). Pour ce faire, 686 syllabes isolées (CV et VC, 7 locuteurs, différentes voyelles) ont été étudiées. C'étaient des traits indépendants, si possible, de la voyelle associée qui étaient recherchés. Différentes grandeurs ont été investiguées : durée de la phase transitoire, positions et mouvements de formants, locus, VOT, répartition de l'énergie

dans le spectre et forme des différentes bosses (pendant le burst, au début et pendant la transition vocalique).

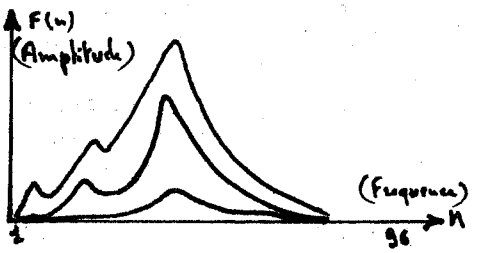
La plupart de ces grandeurs se sont révélées décevantes. En particulier la notion de formant, très utile pour les sons stationnaires et les semi-voyelles, est difficile à étendre aux bursts des consonnes explosives traitées par la cochlée artificielle : ils ont alors très souvent un comportement aléatoire et fugitif.

Les deux grandeurs qui donnent les meilleurs résultats sont la répartition de l'énergie dans le spectre et la forme des différentes bosses pendant le burst, ou s'il est absent au début de la transition. Les résultats qualitatifs obtenus étaient conformes aux règles explicitées au § 4.

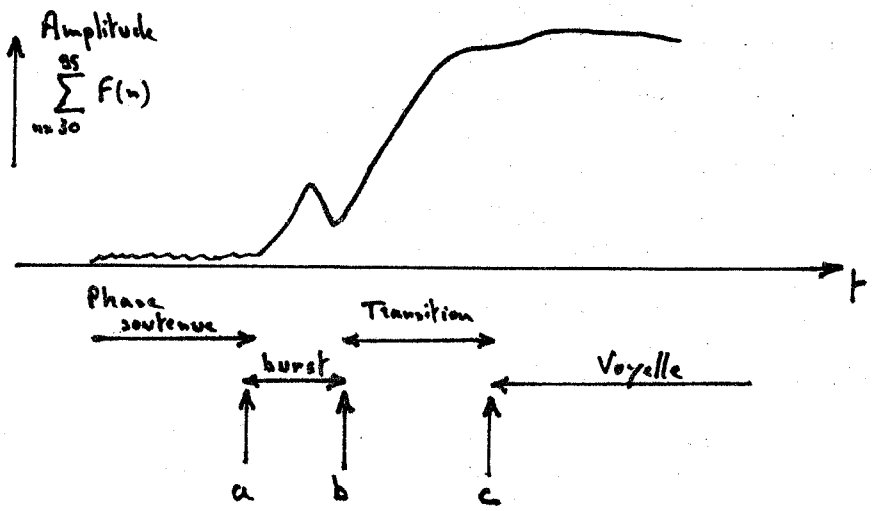
2. - DONNEES UTILISEES ET LOCALISATION.

Les consonnes explosives étudiées étaient prononcées distinctement sous forme de syllabes CV, avec un excellent rapport S/B.
(24 locuteurs + 10 locutrices) x (P, T, K, B, D, G) x (I, É, E, A, O, U, OU,AN,ON) = 1836 syllabes. Le capteur utilisé pour les observer était une cochlée artificielle (Alinat, 1975) fournissant une fonction F(n) toutes les 4 ms.

La localisation de l'instant ou de l'intervalle de temps pendant lequel il faut rechercher les paramètres qui permettront la classification P ou T ou K, a été faite en tenant compte des études antérieures citées au § 1.



Dans le cas des syllabes CV étudiées ici, le déroulement temporel des évènements est :



Dans la présente étude, le point a est défini comme indiquant le début de l'augmentation d'énergie dans la partie haute du spectre (au-delà de n = 30) et le point b comme indiquant le début de l'augmentation d'énergie dans la partie basse du spectre (en-deçà de n = 30). La définition du point C est moins précise : c'est l'instant à partir duquel les formants sont à peu près stables. Il peut arriver, notamment pour /P/ et surtout /B/, que b se produise en même temps ou même avant a (absence de burst). Il peut également arriver, notamment pour /K/ que c soit quasiment simultané de b (absence de transition). En revanche, s'il n'y a ni burst ni transition, il faut considérer qu'il n'y a pas présence d'une consonne explosive (et donc seulement une voyelle dont l'attaque

est plus ou moins explosive).

Il semble que l'intervalle de temps le plus intéressant soit à peu près :

- le premier tiers du burst si ce dernier est présent
- les 10 ms suivant le point a s'il n'y a pas de burst

Dans la présente étude, après quelques essais, il a été décidé d'extraire les paramètres sur la fonction F(n) qui suit a lorsqu'il y a un burst et sur la fonction F(n) au point a s'il n'y a pas de burst.

Notons que le succès de la classification P ou T ou K dépend pour une grande part d'une localisation correcte.

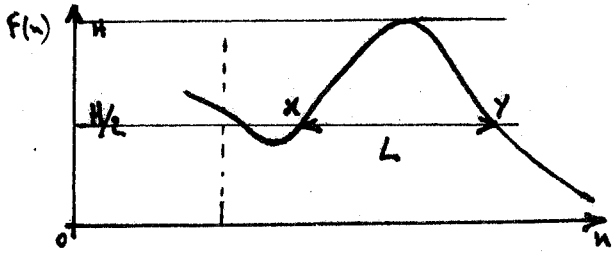
3. - PARAMETRES UTILISES.

D'après toutes les études antérieures, les paramètres suivants ont semblé intéressants et ont été systématiquement extraits pour les 1836 syllabes :

- Durée du burst d = durée ab (unité 4 ms)
- Rapport $\frac{\text{énergie basse fréquence}}{\text{énergie haute fréquence}} = \sum_{30}^{50} F(n) \Big/ \sum_{50}^{95} F(n) = \alpha$

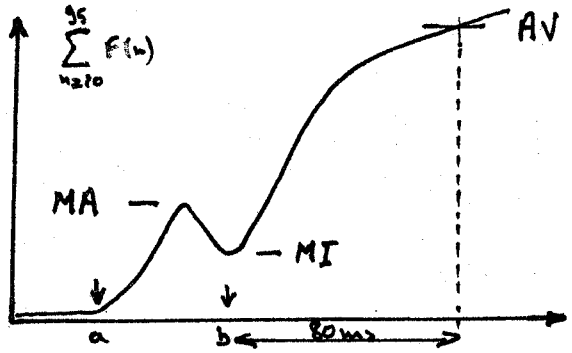
- Position M du plus grand sommet au-delà de n = 30. Lorsque le maximum de F(n) au-delà de n = 30 est situé avant n = 40 et qu'il n'existe pas de sommets encadrés de minimums suffisament bas, on note "BF" au lieu de la position M.

- Largeur L du pic à mi-hauteur du sommet et les positions des points X et Y



- Les rapports $\frac{MA}{AV}$ et $\frac{MA}{MI}$

avec MA, MI et AV définis par la figure ci-contre :
On met 0 s'il n'y a pas de rebond.



- La position du F2 de la voyelle associée.

Les localisations et les extractions de paramètres ont été exécutées par programmes en temps différé. Il apparut très vite que le rapport α était sans intérêt. Son calcul fut donc supprimé. Pour certains locuteurs un réglage trop élevé du zéro de CAD a empêché une bonne numérisation de certains bursts : les syllabes correspondantes ont été supprimées. La détermination de F2 utilisée ici est très grossière (du fait qu'il n'y a pas vraiment localisation de la voyelle) : les plus grosses erreurs ont été corrigées après coup.

4. - ETUDES DES PARAMETRES OBTENUS ET ETABLISSEMENT DES REGLES DE CLASSIFICATION.

Afin d'étudier l'intérêt des divers paramètres des histogrammes ont été établis par type de syllabe pour la durée, la largeur L, F2 - M, MA/AV et MA/MI. En examinant ces histogrammes il est apparu que :

a) Pour la durée :

- Durée P < durée T < durée K
- Durée B < durée D < durée G
- Comportement très différent pour sourdes et sonores
- Dépend peu de la position de F2
- Pas de bonne séparation PB - TD - KG

b) Pour la largeur L :

- Comportement similaire pour sourdes et sonores
- Permet une bonne séparation entre KG d'une part et PB, TD d'autre part et ce, indépendamment de la voyelle associée

c) Pour F2-M :

- Comportement similaire pour sourdes et sonores
- Généralement F2 - Max $\prec$ 0 pour TD
  F2 - Max proche de 0 pour KG
- Dépendant de la position de F2

d) Pour MA/AV et MA/MI :

- Rebonds K plus nombreux que T plus nombreux que P
- Rebonds G plus nombreux que D plus nombreux que B
- Comportement différent pour sourdes et sonores
- Pas de bonne séparation PB - TD - KG

A partir de ces données, des règles de classification ont été établies par l'opérateur par intéraction avec le calculateur en restant soumis aux impératifs suivants :

- Il s'agit d'une classification PB ou TD ou KG, donc commune aux sourdes et aux sonores
- Les règles doivent être les plus indépendantes possible de la voyelle associée
- Les règles doivent être si possible simples.

Seuls L et F2 - M sont indépendants de la nature sourde-sonore. Grosso-modo la classification adoptée consite à faire une première séparation en utilisant L (indépendant de la voyelle associée), une seconde séparation en utilisant F2 - M (dépendant de F2) et enfin à affiner en utilisant accessoirement d et Y.

Finalement les règles utilisées pour la classification sont :

$L \leqslant 24$ | Durée > 1 et $C \leqslant$ F2-M $\leqslant D \rightarrow$ KG
| sinon F2-M $\leqslant E \rightarrow$ TD, sinon $\rightarrow$ PB

$30 > L > 24$ | Durée > 1 et $-6 \leqslant$ F2-M $\leqslant 0 \rightarrow$ KG
| sinon F2-M $\leqslant E \rightarrow$ TD, sinon $\rightarrow$ PB

$30 \leqslant L$ | F2-M $\leqslant E$ ou Y > 39 $\rightarrow$ TD
| sinon $\rightarrow$ PB

et enfin BF $\rightarrow$ PB.

Les limites C D et E dépendent du F2 de la voyelle associée. Pour la distinction entre PB et TD, on a été amené à faire intervenir Y : la raison en est que la position M du sommet, excellent paramètre pour localiser une bosse étroite, n'est pas valable pour localiser une bosse large. Dans ce dernier cas il aurait été préférable d'utiliser une grandeur du type centre de gravité. Conformément à l'expérience de Liberman (1952), KG est donc caractérisé par un burst étroit situé près du F2 de la voyelle associée, ; un burst étroit en fréquence situé au-dessus de F2 correspond à TD et à PB en-dessous. Ce qui n'apparaît pas dans l'expérience de sytnhèse de Liberman, c'est le fait que très souvent TD et PB (et non pas KG) sont caractérisés par un burst large, situé en dessus de F2 pour TD, en-dessous pour PB.

5. - RESULTATS.

Pour l'ensemble des 34 locuteurs, les matrices de confusion suivantes ont été obtenues :

Prononcés	P	T	K		B	D	G
Reconnues P	90 %	6	2	B	95	13	6
T	5	91	5	D	2	82	3
K	3	2	92	G	2	3	90

Les différences entre homme et femme ne sont pas plus marquées qu'entre sourdes et sonores.

Ces résultats peuvent être comparés à ceux relatifs à la classification par des auditeurs humains dans diverses conditions : G.A. Miller (1955), H. Winitz (1972), C. La Rivière (1975).

Notons que la présente étude était relative aux consonnes explosives en position CV. Il serait intéressant de l'étendre aux syllabes VC et au cas où le phonème associé est une consonne fricative ou une semi-voyelle. En particulier que deviennent alors les règles qui faisaient intervenir F2 ? Notons enfin que la présente étude procédait par analyse : il serait intéressant de la contrôler par une étude procédant par synthèse.

REFERENCES

Alinat P. 1975 "Etude des phonèmes de la langue française au moyen d'une cochlée artificielle. Application à la reconnaissance de la parole" Rev. Tech. Thomson-CSF vol 7 n° 1 pp. 91 à 123

Halle M., Hughes G.W., Radley J.P.A, 1957, "Acoustic properties of stop consonants" J.A.S.A. vol 29 n° 1, pp. 107-116

La Rivière C., Winitz H., Herriman D, 1975 "Vocalic transitions in the perception of voiceless Initial stops "J.A.S.A. n° 2 pp.470-475

Liberman A.M., Delattre P., Cooper F.S., 1952 "The role of selected stimulus variables in the perception of the unvoiced stop consonants" Amer. Journ. of Psychol. vol LXV pp. 497-516

Miller G.A., Nicely P.E., 1955 "An analysis of perceptual confusions among some english consonants" J.A.S.A. vol 27 pp. 338-352

Winitz H., Scheib M.E., Reeds J.A., 1972, "Identification of stops and vowels for the burst portion of /p,t,k/ isolated from conversationnal speech" J.A.S.A. vol 51 pp.1309 -1317.

9èmes JOURNEES D'ETUDE SUR LA PAROLE

LANNION 31 mai - 2 juin 1978

DETECTION AUTOMATIQUE DE TRAITS ACOUSTIQUES, EN TEMPS REEL, DANS LA PAROLE CONTINUE

M. LAMOTTE, H. LEM, C. VIGNERON, M.-J. VIGNERON

Laboratoire d'Electricité et d'Automatique
Université de NANCY 1

.RESUME

La finalité du système décrit est de reconnaître automatiquement, et en temps réel, des mots dans la parole continue, ceci à l'aide d'un matériel très restreint (un seul microprocesseur).

Après une analyse du signal vocal classique, mais assez rudimentaire (analyseur 8 canaux), la parole est segmentée en zones soutenues et transitoires. Une reconnaissance échantillon par échantillon est alors effectuée à l'aide d'une méthode originale, utilisant des cellules à tests constituant une "machine" de traitement parallèle. Un arrangement des échantillons reconnus élabore une chaîne de caractères.

La reconnaissance au niveau mot est seulement en cours de mise au point.

DETECTION AUTOMATIQUE DE TRAITS ACOUSTIQUES, EN TEMPS REEL, DANS LA PAROLE CONTINUE

M. LAMOTTE, H. LEM, C. VIGNERON, M.-J. VIGNERON

Laboratoire d'Electricité et d'Automatique
Université de NANCY 1

.**SUMMARY**

The purpose of this paper is to describe an automatic procedure for real time recognition of words in continuous speech. The needed equipment is restricted to only one microprocessor. In a first stage, the vocal signal is analyzed through a classical but sketshy method using a eight channel analyzer ; then speech is segmented into sustained and transition zones. Using testing cells operating as a parallel treatment machine, a new method is proposed for sample for sample recognition. A character string is obtained from recognised samples.

Real work recognition is in progress.

INTRODUCTION

Quelques rares machines de reconnaissance de la parole, servant à la commande verbale de processus divers, sont actuellement commercialisées. En général, il s'agit de configurations, tant en matériel qu'en informatique, assez lourdes et assez coûteuses (de l'ordre de 20 000 dollards). Les systèmes proposés par la Threshold Technology Inc. (TTI), et la Scope Electronics Inc. aux USA, avec ceux vendus par la NIPPON Electronic Company (NEC) au Japon, sont à notre connaissance les seuls sur le marché actuel.

En France, le CNET de Lannion a réalisé une machine, "CHARLES" très intéressante, qui reconnaît des mots isolés en langage CHOKINA.

Notre étude a consisté à réaliser entièrement une petite machine peu coûteuse, reconnaissant en temps réel des mots dans la parole continue.

ANALYSE ET PRETRAITEMENT

L'analyseur réalisé est des plus classiques dans son principe. Il est du type "Vocoder" à canaux très simplifié, et possède la particularité de ne comporter que huit voies.

Le signal vocal traverse un banc de filtres. Les signaux obtenus sont ensuite redressés et intégrés durant une fenêtre temporelle de 10ms.

Un multiplexage et une conversion analogique numérique terminent cette chaîne d'acquisition.

SEGMENTATION DE LA PAROLE

Il est à remarquer que les traitements effectués après l'analyse sont réalisés grâce à un microprocesseur 2650 de la firme "Signetics", travaillant avec un "bus" de données de 8 bits, un jeu de 75 instructions et un temps de cycle de base de 3µs. La mémoire utilisable a une taille de 5 kilooctets.

Nous avons choisi un type de segmentation de la parole, séparant les zones soutenues des zones transitoires. Par la suite, la reconnaissance s'effectuera sur les soutenues.

Contrairement à ce qui se pratique couramment, nous n'opérons pas de normalisation, ceci pour deux raisons : le microprocesseur ne possédant pas de division câblée ne nous permettait pas d'implanter ce traitement si on voulait conserver le temps réel pour la reconnaissance. En outre, il nous a paru intéressant de garder la dynamique du signal comme paramètre prosodique.

L'algorithme de segmentation comporte deux paramètres : le premier tient compte des variations d'énergie dues au passage d'un échantillon à un autre (H. LEM, Thèse de Spécialité, 1976). Soit $\underline{X}(t)$ le vecteur échantillon dont les 8 composantes $x_i(t)$ sont les valeurs numériques des énergies délivrées par chaque canal à l'instant t.

L'expression du premier paramètre est :

$$VE(t) = \sum_{i=1}^{8} |x_i(t) - x_i(t-1)| \ .$$

Les soutenus seront caractérisés par une faible variation d'énergie, alors que les transitoires sont le siège d'une brusque et importante activité. La détection des points de segmentation est réalisée à l'aide d'un deuxième paramètre (ΔVE), tenant compte de la rapidité de variation d'énergie.

Le programme [SEG] traite principalement l'algorithme permettant de trouver les points de segmentation. Il se divise en deux branches, séparant les phonèmes en deux classes : les voyelles et les consonnes à forte énergie d'une part, et les consonnes à faible énergie, ainsi que les silences, d'autre part. En outre, [SEG] est considéré comme programme principal de tout le système et appelle tous les sous-programmes.

Le déroulement des opérations apparaît sur l'organigramme de la figure 1 .

RECONNAISSANCE ET TRAITEMENT DES ECHANTILLONS, AFIN DE GENERER UNE CHAINE DE CARACTERES

Il s'agit, à ce niveau, d'étiqueter chaque échantillon afin de savoir à quelle classe de phonèmes il appartient. En conséquence, nous avons mis en oeuvre une méthode tenant compte des traits phonético-acoustiques de la parole. Nous avons travaillé sur 28 classes phonémiques. En effet, les distinctions de type [ɑ, a] ou [o, ɔ] n'ont pas été faites. D'autre part, nous n'avons pas répertorié les semi-voyelles.

La méthode utilisée consiste à présenter un échantillon de parole en parallèle à huit tests logiques (réponse par oui ou non) constitués en cellule. La combinaison des réponses aux tests donne un mot de huit bits (code) représentatif de l'échantillon étudié.

Les tests utilisés sont très simples (comparaisons, additions, soustractions) et ne demandent que très peu de temps calcul.

D'autre part, la méthode est souple. Dans une phase de mise au point, on peut remplacer un test par un autre sans changer le reste de la structure. Il est à remarquer enfin qu'une telle cellule peut délivrer théoriquement 256 codes différents.

De nombreux essais ont montré qu'une seule cellule n'était pas suffisante. Nous sommes arrivés à la configuration représentée sur la figure 2, où trois cellules ont été mises en cascade et complétée par des tests spécialisés pour les consonnes à faible énergie.

La réponse de la machine est donc un code sur huit bits, correspondant à un phonème ou une confusion de phonèmes. Nous avons relevé la fréquence d'apparition des codes afin de repérer les plus probables.

Le dépouillement des résultats nous a donné 15 phénomènes seuls bien triés, et 21 confusions bien séparées entre elles également. En général, les consonnes sont mieux discriminées que les voyelles : s, f, z, v sont séparés ; [ch,j] , [p,t,k] , [b,d,g] restent en confusion. a, o, ou, u, e, è, é, in, un sont triés ; par contre, on les retrouve parfois dans les confusions restantes. A ce niveau, un chiffrage de reconnaissance est prématuré actuellement. De nombreux essais doivent encore être faits.

L'identification du soutenu générant un caractère consiste à retenir le ou les codes correspondant aux échantillons qui sont les plus nombreux durant le soutenu considéré.

Après quelques tests corrigeant les erreurs restantes de la segmentation (soutenu aberrant, trop long, ou trop court), un regroupement des codes différents d'un même phonème est effectué sous un même numéro (code ASCII). Le nom du soutenu sera celui du phonème ayant la plus grande fréquence d'apparition.

Si d'autres phonèmes apparaissent avec la même fréquence F ou F-1, le soutenu sera étiqueté par les numéros (ASCII) de ces phonèmes, tout en se limitant à 6 confusions.

Les premiers résultats obtenus permettent déjà de faire quelques remarques : on constate l'existence de répétitions dues, d'une part au fait qu'un soutenu trop long est arbitrairement coupé en deux, d'autre part à une mauvaise prononciation pouvant provoquer un point de segmentation parasite. Quelques rares omissions sont également observées, ainsi que quelques erreurs de reconnaissance.

Un programme de reconnaissance des mots dans la parole continue, utilisant un niveau lexical, est en cours de mise au point.

CONCLUSION

La qualité relativement bonne de la chaine obtenue (en se basant sur les premiers résultats) permet de penser qu'avec l'adjonction d'un niveau lexical tenant compte des erreurs restantes, nous obtiendrons une reconnaissance correcte sur une vingtaine de mots de commande. Il est évident que la recherche devra se poursuivre et que quelques améliorations peuvent être apportées aux différents niveaux.

REFERENCES

LEM, H. , 1976, Thèse de Spécialité.

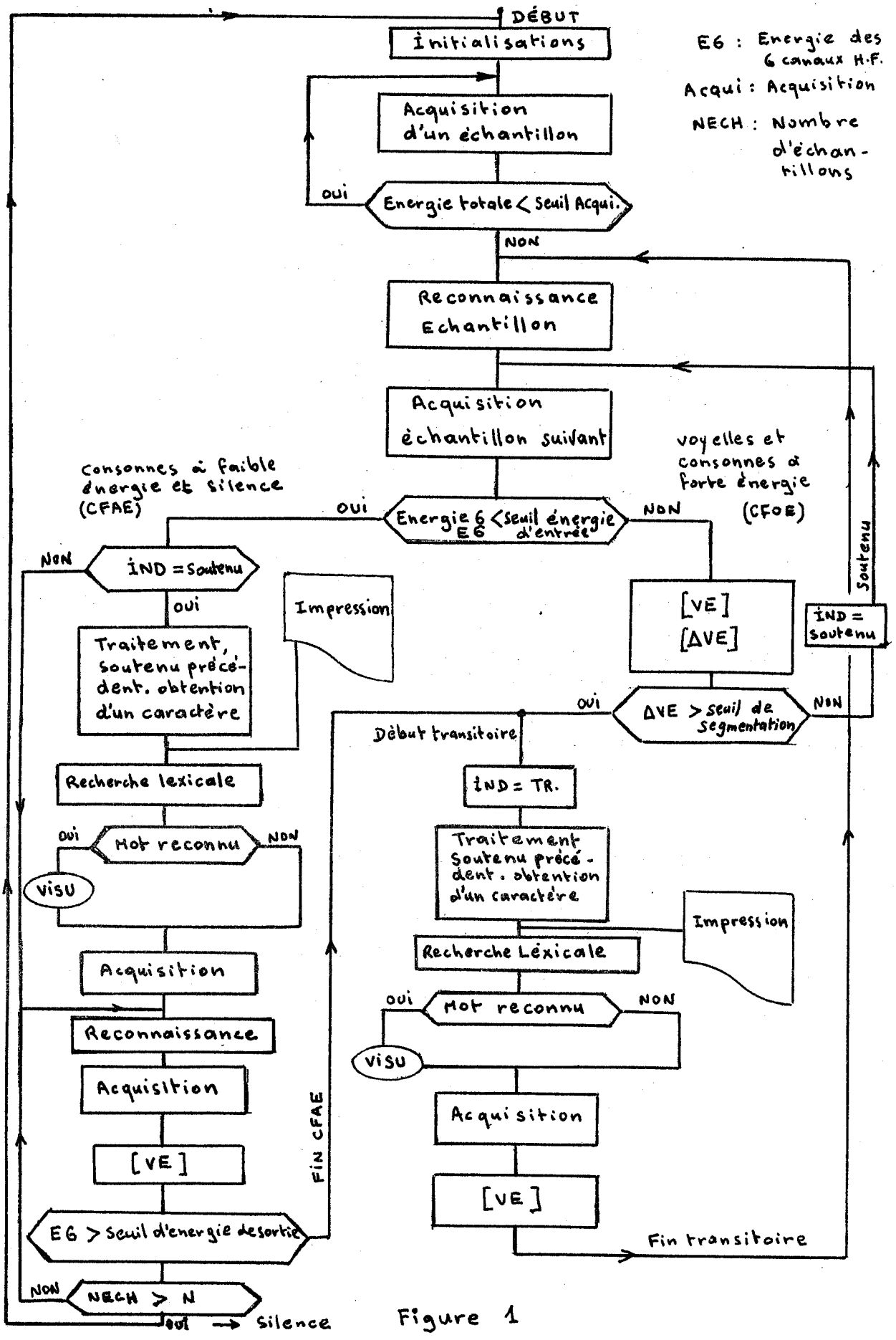

DÉBUT
Initialisations
Acquisition d'un échantillon
oui
Energie totale < Seuil Acqui.
NON
Reconnaissance Echantillon
Acquisition échantillon suivant
Consonnes à faible énergie et silence (CFAE)
oui
Energie E6 < Seuil énergie d'entrée
NON
voyelles et consonnes à forte énergie (CFOE)
NON
IND = Soutenu
oui
Impression
Traitement, soutenu précédent. obtention d'un caractère
Recherche lexicale
oui
Mot reconnu
NON
VISU
Acquisition
Reconnaissance
Acquisition
[VE]
E6 > Seuil d'energie de sortie
NON
NECH > N
oui
Silence
FIN CFAE
[VE]
[ΔVE]
ΔVE > seuil de segmentation
oui
NON
Début transitoire
IND = TR.
Traitement Soutenu précédent. obtention d'un caractère
Impression
Recherche Lexicale
oui
Mot reconnu
NON
VISU
Acquisition
[VE]
Fin transitoire
Soutenu
IND = soutenu
E6 : Energie des 6 canaux H.F.
Acqui : Acquisition
NECH : Nombre d'échantillons

Figure 1

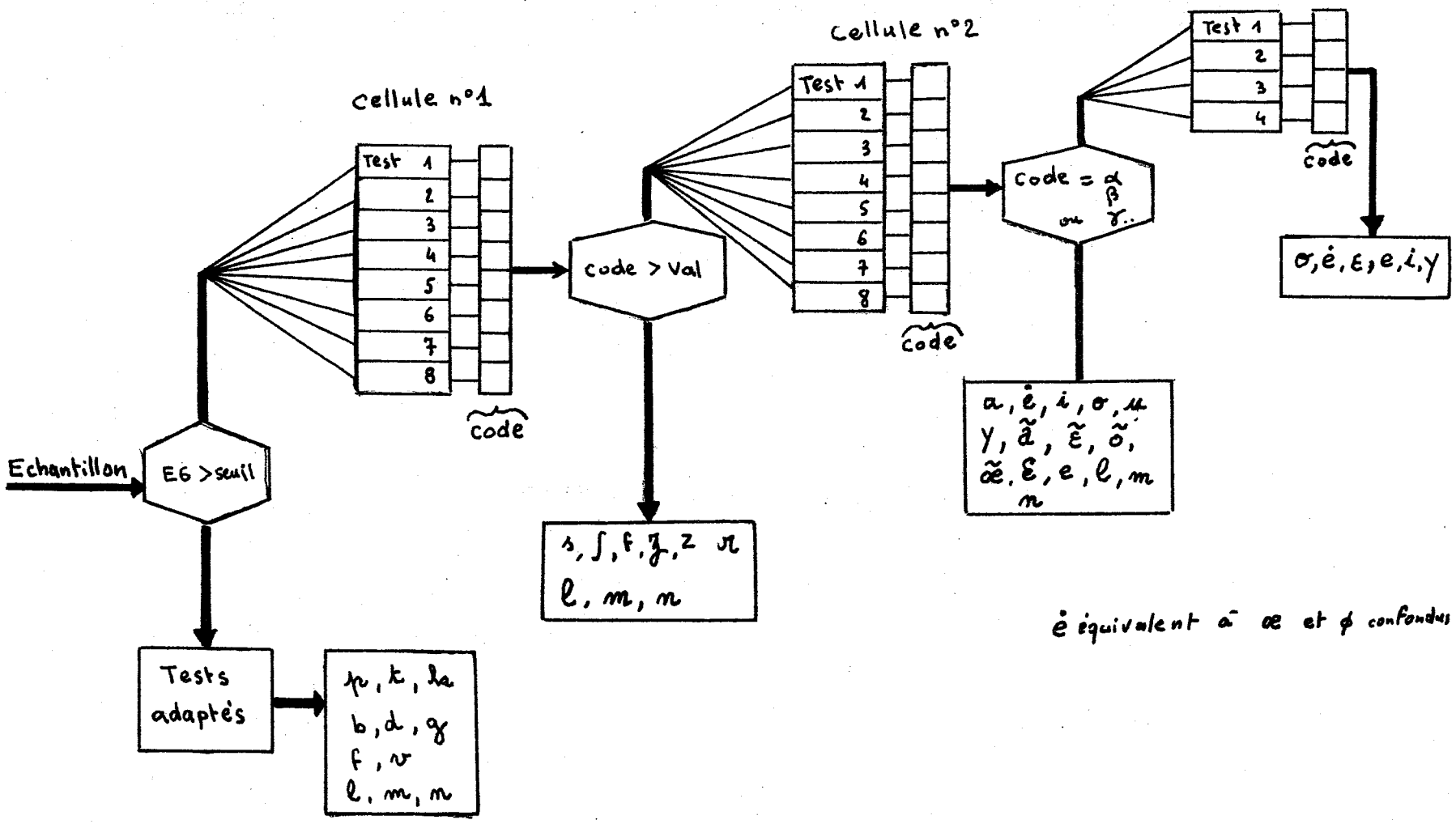

Cellule n°1
Test 1
2
3
4
5
6
7
8
code
Cellule n°2
Test 1
2
3
4
5
6
7
8
code
Test 1
2
3
4
code
Echantillon
EG > seuil
code > Val
Code = α β ou γ..
Tests adaptés
s, ʃ, f, ʒ, z ɾ
l, m, n
p, t, k
b, d, g
f, v
l, m, n
σ, ė, ɛ, e, i, y
a, ė, i, σ, u
y, ã, ɛ̃, õ,
œ̃, ɛ, e, l, m
n
è équivalent à œ et ɸ confondus

Figure 2

9èmes JOURNEES D'ETUDE SUR LA PAROLE

LANNION 31 mai - 2 juin 1978

UTILISATION DE METHODES SYNTAXIQUES POUR LA DETECTION AUTOMATIQUE DES TRAITS PHONETIQUES EN RECONNAISSANCE DE LA PAROLE

M. BAUDRY - B. DUPEYRAT

CEN/SACLAY
Service d'Electronique de Saclay
Service d'Instrumentation pour la Recherche - Bât 28 - pièce 111B
Boite postale n° 2
91190 GIF-sur-YVETTE

.RESUME

Le système de reconnaissance utilisé travaille sur le signal acoustique direct. Le calculateur utilise les intervalles entre extrema et les amplitudes des extrema du signal.

Les traits phonétiques extraits par le système sont les suivants :

1 -	Voisement	Non voisement	Occlusion
2 -			voisée, non voisée
3 -	friction, Nasalité, liquide, voyelle		

La mélodie est calculée, mais n'est pas encore utilisée pour la reconnaissance.

Pour extraire ces informations, on utilise des règles de récriture à trois niveaux du traitement.

Ces règles portent :

1 - sur la statistique des passages par zéro pour la décision voisée/non voisée,
2 - sur l'ordre d'apparition, les intervalles et les amplitudes des extrema, pour la détection du début du cycle de voisement,
3 - sur l'enveloppe du signal dans les zones voisées pour la détection des traits de friction, de nasalité, des liquides et des voyelles.

Résultat : l'application de ces méthodes conduit à une segmentation sûre du message acoustique qui donne avant la phase de reconnaissance proprement dite des informations sur la nature phonétique des segments.

AUTOMATIC DETECTION OF PHONETIC FEATURES IN SPEECH RECOGNITION BY SYNTACTIC METHODS

M. BAUDRY - B. DUPEYRAT

CEN/SACLAY
SES/SIR - Bât 28 - pièce 111B
B.P n° 2
91190 GIF SUR YVETTE

. SUMMARY

The used pattern recognition system works on the direct acoustic signal described by the extrema amplitudes of the signal and the intervals between them.

The output phonetic features which are obtained by the system are :

1 -	Voicing	unvoicing	occlusion
2 -			voiced, unvoiced
3 -	fricative, nasal, liquid, vowel		

The melody is computed but still unused for recognition. Rewriting rules are used at three levels of the process in order to extract the informations described above.

The rules operate on :

1 - the zero-crossings statistic for the voiced/unvoiced decision,
2 - the order of occurrences, the intervals and the amplitudes of the extrema for detecting the beginning of the pitch,
3 - the signal envelop in the voiced zones for detecting the fricatives, the nasals, the liquids and the vowels.

RESULTS : These methods, when applied, yield a safe segmentation of the acoustic message which gives informations about the phonetic character of the segments, before the recognition stage.

INTRODUCTION

Le signal est analysé directement dans sa représentation amplitude-temps. Les seules informations transmises en calculateur sont la position et l'amplitude des extremum. L'échantillonnage est effectué à 10 KHz et les amplitudes sont codées sur 64 niveaux. Les programmes fonctionnent, en temps réel, sur mini-ordinateur MULTI-20 de 32 K (M. BAUDRY, B. DUPEYRAT, 1977).

DETECTION DES TRAITS PHONETIQUES

L'application des méthodes syntaxiques (développées plus loin) permet la détection des traits phonétiques suivants :

- les silences occlusifs (détection précise et analyse des bruits de friction et d'explosion présents),
- le caractère fricatif ou voisé du signal ainsi que le temps d'établissement du voisement (VOT),
- la micromélodie et la mélodie de la parole analysée par la détection précise de chaque début de cycle de voisement (DCV),
- l'opposition voyelle consonne (en suivant l'évolution de l'amplitude maximale de chaque DCV).

CARACTERISTIQUES DES GRAMMAIRES UTILISEES (K.S.FU - 1974, DE MORI - 1976)

Les éléments du langage ou formes élémentaires possèdent un ensemble d'attributs (valeurs numériques : amplitudes, durées, etc.). Les propriétés de la forme changent suivant les valeurs de ces attributs. Nous pouvons donc les considérer comme une information sémantique associée à la forme élémentaire.

Les règles de récriture pourront donc être validées par la vérification de prédicats utilisant les attributs. Signalons d'abord que l'analyse est faite de bas en haut.

Le traitement en temps réel du signal nous amène :

- d'une part, à faire l'analyse grammaticale de gauche à droite, c'est à dire à chaque acquisition d'un nouvel élément terminal,
- d'autre part, à construire un arbre syntaxique non ambigu. Pour cela, la grammaire est construite de telle façon que, les règles étant classées, on utilise systèmatiquement la première règle qui s'applique. Cela revient pratiquement à construire l'arbre en montant le moins vite possible vers l'axiome.

REMARQUE : Les grammaires ont été construites manuellement.

EXEMPLES

1°) Recherche du temps d'établissement du voisement et détection des cycles de voisement dans une plosive

Nous appelons AP et AN respectivement une arche positive et négative définie par les passages par zéro du signal (figure 1). Chaque arche possède deux attributs :

. sa durée Δt
. son amplitude maximale α

La recherche des DCV, en début de parole, est rendue délicate par les variations rapides des cycles de voisement et par les perturbations apportées par les plosives. La figure 1 montre l'exemple de son /k $\tilde{\varepsilon}$/ (mot kin). L'arche AN_6 est due à l'explosion, elle est précédée par un bruit de friction ; elle ne doit pas être prise comme DCV sinon la distance entre les arches AN_6 et AN_9 donnerait une valeur erronée de la période fondamentale.

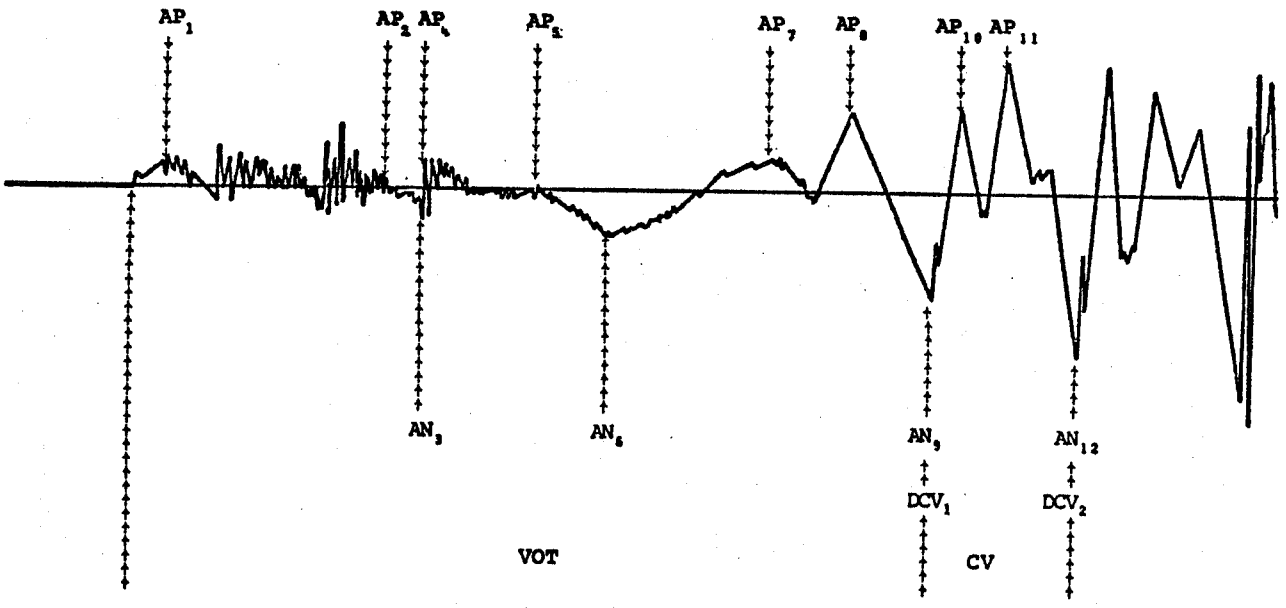


FIGURE 1 - son /k $\tilde{\varepsilon}$ /

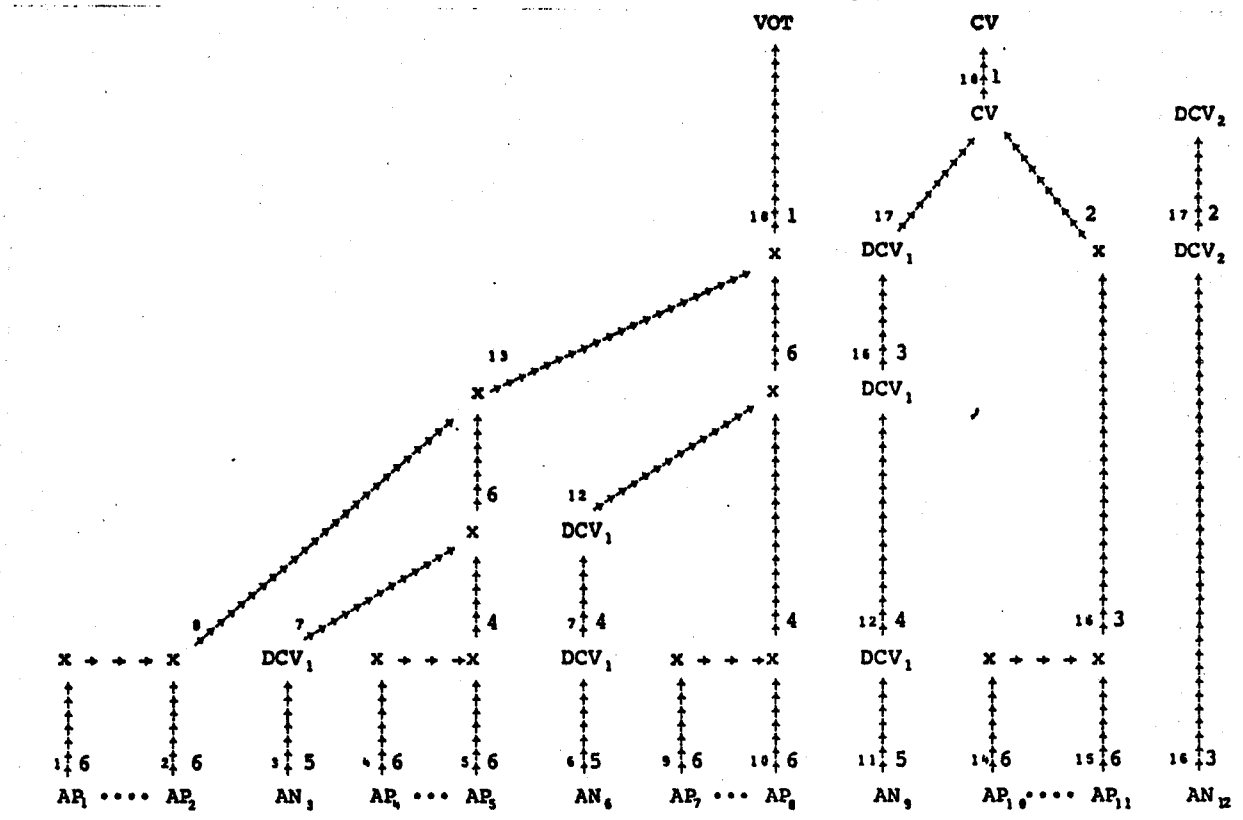


FIGURE 2 - Arbre syntaxique du son /k $\tilde{\varepsilon}$/

Les règles de récriture sont appliquées aux vocabulaires terminal VT_3 et non terminal VN_3.
La condition C_1 utilise des seuils relatifs à Δt et α.
Le prédicat P_1 utilise des relations entre les Δt et α.
X est une chaîne quelconque d'arches positives et négatives ne répondant pas à la condition C.
La règle 4 correspond à l'abandon d'une arche AN de DCV possible pour une nouvelle arche AN.
L'initialisation est terminée lorsque la règle 3 est utilisée.
Si cette analyse échoue, nous faisons l'hypothèse que le segment de parole analysé n'est pas vraiment voisé, la décision voisée-non voisée est réexaminée.
VOT correspond au temps d'établissement du voisement.

VT_3 = {AN, AP}
VN_3 = {VOT, DCV_1, DCV_2, X}

1 VOT,CV → X,CV

2 CV,DCV_2 → DCV_1, X, DCV_2

3 DCV_1, X, DCV_2 → (DCV_1, X, AN[C_1]) [P_1 (DCV_1, AN)]

4 X, DCV_1 → DCV_1, X, DCV_1

5 DCV_1 → AN [C_1]

6 X → AN [¬ C_1] | AP| X, AP| X, X

La figure 2 montre la construction de l'arbre syntaxique correspondant au signal de la figure 1. A gauche des flèches figure le numéro d'ordre dans lequel les règles sont utilisées et à droite le numéro de la règle.

2°) Segmentation voyelle-consonne

La figure 3 représente l'enveloppe obtenue en conservant pour chaque cycle de voisement (CV) le point d'amplitude maximal α.

On construit à partir des éléments terminaux CV d'attributs α des segments :

- d'amplitude montante LM
- d'amplitude stable LH
- d'amplitude descendante LD

VT_5 = (CV)

VN_5 = (LM, LH, LD, L)

L'élément L permet d'identifier des cas de segmentations particuliers.

1. $PV \rightarrow LM|LH|LD|PV,LM|PV,LH|PV,LD$

2. $L_1 \rightarrow CV_1$

3. $LH_k \rightarrow L_k, CV_{k+1} \quad \{P_s\}$
4. $LH_i \rightarrow LH_i, CV_k \quad \{P_s\}$
5. $LH_i, L_k \rightarrow LH_i, CV_k \quad \{P_{lm} \vee P_{ld}\}$
6. $LH_i, LM_k \rightarrow LH_i, CV_k \quad \{P_{rm}\}$
7. $LH_i, LD_k \rightarrow LH_i, CV_k \quad \{P_{rd}\}$

8. $LM_k \rightarrow L_k, CV_{k+1} \quad \{P_{rm} \vee P_{lm}\}$
9. $LM_i \rightarrow LM_i, CV_k \quad \{P_{rm} \vee P_{lm}\}$
10. $LM_i, L_k \rightarrow LM_i, CV_k \quad \{P_s\}$
11. $LM_i, LD_k \rightarrow LM_i, CV_k \quad \{P_{rd}\}$
12. $LM_i \rightarrow LM_i, L_k, CV_{k+1} \quad \{P_{rm}(L_k, CV_{k+1}) VP_{lm}(LM_i, CV_{k+1})\}$

13. $LD_k \rightarrow L_k, CV_{k+1} \quad \{P_{rd} \vee P_{ld}\}$
14. $LD_i \rightarrow LD_i, CV_k \quad \{P_{rd} \vee P_{ld}\}$
15. $LD_i, L_k \rightarrow LD_i, CV_k \quad \{P_s\}$
16. $LD_i, LM_k \rightarrow LD_i, CV_k \quad \{P_{rm}\}$
17. $LD_i \rightarrow LD_i, L_k, CV_{k+1} \quad \{P_{rd}(L_k, CV_{k+1}) VP_{ld}(LD_i, CV_{k+1})\}$

La règle 2 sert uniquement à initialiser l'analyse et n'est utilisée qu'une fois pour chaque PV.

Les règles (3), (6, 8, 16) et (7, 13, 11) définissent le début des lignes LH, LM et LD (respectivement).

Les règles 4, 9 et 14 définissent la continuité normale des lignes LH, LM et LD (respectivement).

Les règles (5, 6, 7), (10, 11) et (15, 16) définissent la fin des lignes LH, LM et LD (respectivement).

Les règles 12 et 17 (après 10 et 15) annulent la transition et assurent la continuité des lignes LM et LD (cas de deux points stables pendant une montée ou descente).

CONCLUSION

Ces méthodes syntaxiques permettent une détection rapide et sûre des indices phonétiques cités plus haut pour différents locuteurs. Elles formalisent de façon concise les analyses du signal direct faites visuellement.

Les résultats sont completés par une analyse phonémique. Nous n'avons envisagé pour l'instant que l'application à un dictionnaire de mot isolés.

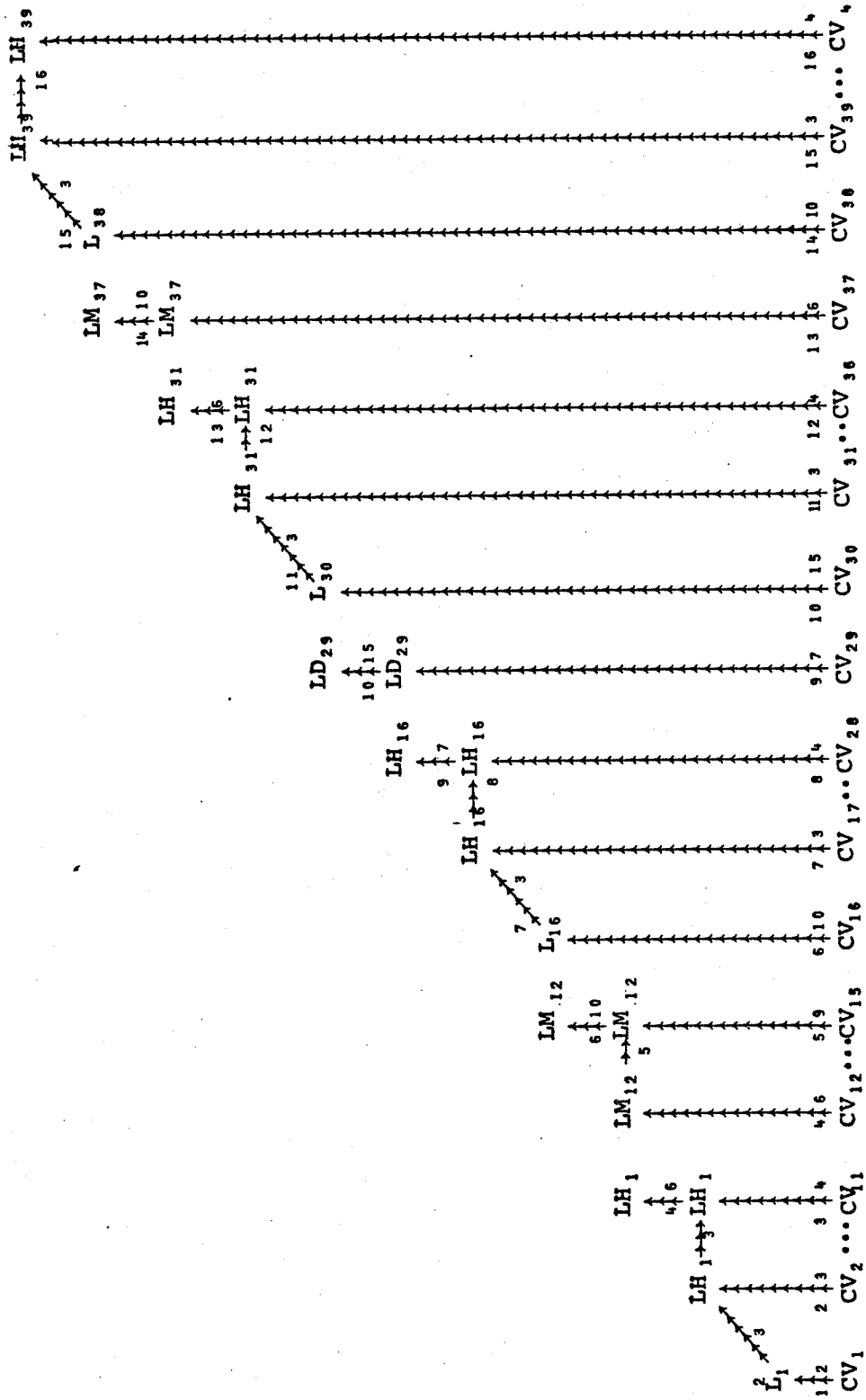


FIGURE 4 - Arbre syntaxique pour la segmentation du mot "voilà"

Signalons que des méthodes syntaxiques sont également développées pour la recherche en dictionnaire. Celle-ci comporte une phase d'apprentissage automatique au cours de laquelle les règles sont créées et classées. Ce travail est en cours d'étude.

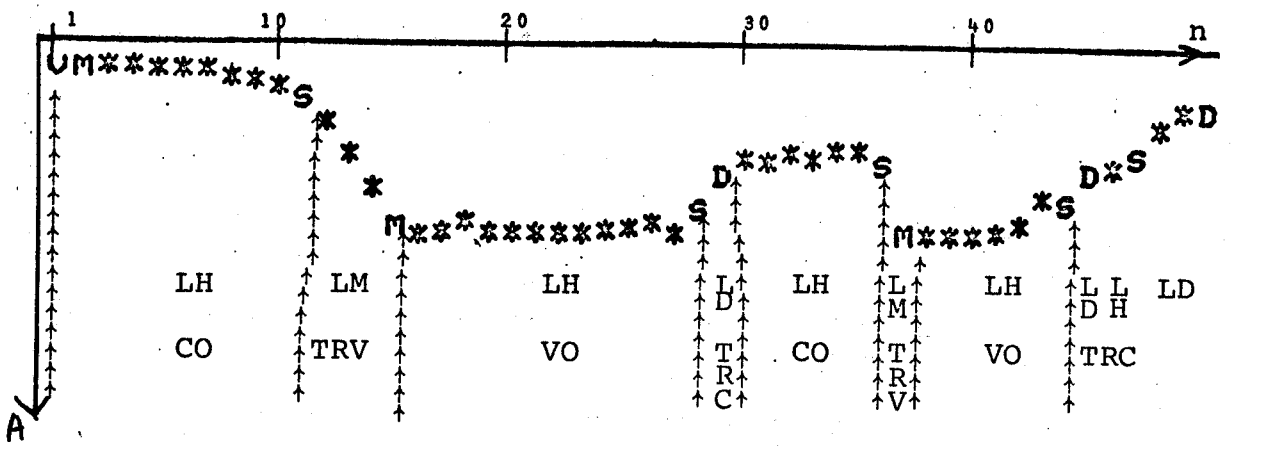


FIGURE 4 - Segmentation du mot "voilà"

REFERENCES

BAUDRY M. , DUPEYRAT B. , 1977 , Analyse du signal vocal en temps réel sur mini-ordinateur. 9ème I.C.A. MADRID (I44)

FU K.S. , 1974, Syntactic methods in Pattern Recognition, Academic Press

DE MORI R. , LAFACE P. , PICCOLO E. , October 1976 , Automatic Dectection and Description of Syllabic Features in Continuous Speech
IEEE Transactions on acoustics - Vol : ASSP 24 n° 5

9èmes JOURNEES D'ETUDE SUR LA PAROLE

LANNION 31 mai - 2 juin 1978

EVALUATION DES INDICES ACOUSTIQUES UTILISES DANS L'ANALYSEUR PHONETIQUE DU SYSTEME KEAL

G. MERCIER CNET - ROUTE DE TREGASTEL - 22301 LANNION

.**RESUME**

Dans cet article, on présente les indices acoustiques utilisés dans le module Soniadel (analyseur phonétique) du système Kéal et permettant

1 - de segmenter le signal en syllabes et phonèmes,

2 - de reconnaître les classes telles que voyelles, consonnes, constrictives, voisées, sourdes, occlusives, nasales, liquides,

3 - d'identifier les phonèmes à l'intérieur de ces classes.

On essaiera d'évaluer à chaque niveau de segmentation ou d'identification l'ensemble des indices utilisés.

SONIADEL = phonétique en breton

SUMMARY

This paper describes in detail the acoustic parameters used by the phonetic analyzer (Soniadel) of the complete speech understanding system "Keal". The system accepts parameters derived from a channel vocoder together with a pitch detector. The frequency bands of this vocoder are given in Fig. 1.

In Fig. 2 and 3, two examples of the outputs given by the phonetic analyzer for two different sentences are shown. In Fig. 3, the intermediary classes of each segment (voiced, unvoiced, plosives, fricatives...) determined by the program are also shown. The identity and the content of each class of phonemes are given in Fig. 5.

The general organization of the phonetic analyzer is presented in section II and summarized in Fig. 4. The main acoustic parameters used to distinguish between the features are listed in section III and in the appendix.

Furthermore in section IV, each step of the program is described in detail : noise-speech separation, segmentation into syllables, segmentation into phonemes, vowel-consonant discrimination, detection of the consonant-class of the phoneme (plosive, fricative, liquid, nasal) detection of the absence of consonant versus the presence of a plosive, in the beginning of a word. The parameters used in each step are detailed and evaluated. Results are presented from figures 6 to 11.

Fig. 12 summarizes the global percentage of the phoneme recognition for four male speakers, the first one being the learning speaker. This table shows the percentage of the number of times that the correct phoneme is within the top m choices out of 31 possible phoneme categories, for m = 1, 2 or 3. The following line of the table indicates the percentage of the number of times that the correct phoneme is within the answers given by the analyzer. The number of such answers depends upon how many candidate phonemes have a score that exceeds a fixed threshold.

The percentage of recognition depends on many factors, the most important ones being : the speaker, the length of the sentences (in words and in time) and the application (the set of sentences). System sensitivity to such factors implies that the results must be interpreted with great caution.

I - INTRODUCTION

L'analyseur phonétique "Soniadel" dont nous esquissons les principales caractéristiques, s'intègre dans le système "Keal" de dialogue homme-machine. Ce système expérimental possède une architecture assez souple pour lui permettre d'utiliser différentes sources d'informations. Celles-ci lui sont fournies par l'intermédiaire de divers modules d'analyse acoustique, phonétique, lexical, syntaxique et de dialogue.

Le rôle de "soniadel" est de représenter sous forme d'une suite de phonèmes le signal de parole s (t) (figures 2 et 3). Celui-ci a été préalablement numérisé par un système d'analyse acoustique dont la fonction est de représenter le signal le plus fidèlement possible mais aussi de réduire le débit.

L'analyseur acoustique utilisé dans les expériences suivantes est un vocodeur à canaux de 14 filtres dont les bandes passantes sont indiquées sur la figure 1. Il est certain que les limites en résolution du vocodeur à canaux conditionnent les résultats de l'analyse phonétique notamment par ses limites en fréquence (300-4200 Hz) et son manque de précision dans les basses fréquences ; nous espérons combler cette lacune prochainement en utilisant d'autres types d'analyse.

CANAL	BANDE PASSANTE
1	250 - 450 HZ
2	450 - 650 HZ
3	650 - 850 HZ
4	850 - 1050 HZ
5	1050 -1300 HZ
6	1300 - 1600 HZ
7	1600 - 1900 HZ
8	1900 - 2200 HZ
9	2200 - 2500 HZ
10	2500 - 2800 HZ
11	2800 - 3100 HZ
12	3100 - 3400 HZ
13	3400 - 3800 HZ
14	3800 - 4200 HZ

FIG.1: CARACTERISTIQUES DU VOCODEUR A CANAUX
FREQUENCY - BANDS OF THE CHANNEL VOCODER

.../...

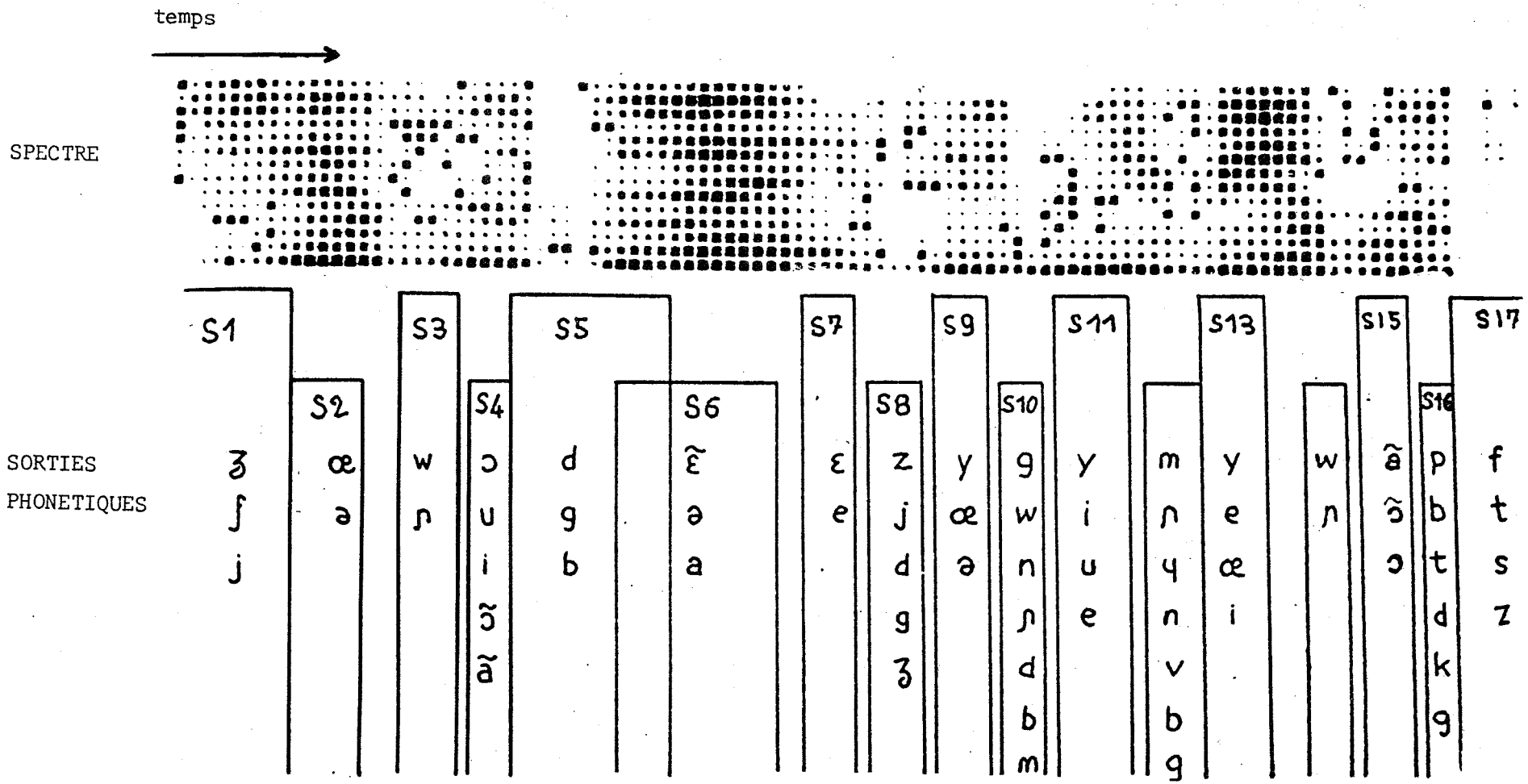


Figure 2 : Résultat de l'analyse phonétique du signal numérisé pour le début de la phrase "Je voudrais le numéro de téléphone de...". Chaque segment est étiqueté par une liste de phonèmes les plus probables.

output of the channel vocoder (spectrum)

Output of the phonetic analyzer for the sentence:"ʒøvudrɛlənumerodt..."

Nous pensons cependant que l'essentiel de l'information du signal de parole se retrouve à la sortie du vocodeur et que les principes et les méthodes de recherche d'indices acoustiques que nous présentons restent valables pour d'autres systèmes d'analyse acoustique.

II - ORGANISATION GENERALE DU MODULE D'ANALYSE PHONETIQUE

Sachant que la parole est constituée d'unités linguistiques de base telles que mots, syllabes, phonèmes, nous allons chercher à localiser et à identifier ces unités.

Nous ne parlerons pas ici de la localisation des mots à l'aide de l'analyse prosodique (J. VAISSIERE, MADRID 1977) ni de l'identification de ces mots (R. VIVES 1976) ; nous nous situons volontairement aux niveaux de la syllabe et des phonèmes.

Etant donné la nature et la complexité du signal de parole et la difficulté du problème, la transcription phonétique n'est pas unique et pour chaque segment localisé le programme de transcription peut fournir plusieurs réponses se présentant dans un ordre de probabilité décroissante (figure 2).

Trois étapes permettent de réaliser cette segmentation :
(figure 4)

- Extraction de paramètres

- Segmentation et localisation des unités de base

- Identification

Le problème n'est cependant pas si simple et ces 3 étapes sont imbriquées les unes dans les autres, en effet, si l'extraction de certains paramètres permet de segmenter et de découvrir la classe de ce segment (fricative, plosive, voisée, sourde...), la connaissance de celle-ci permet à son tour de diriger la recherche vers tel ou tel indice acoustique qui permettra une identification plus précise du segment suivant ou du segment précédent et éventuellement une meilleure segmentation, d'où la nécessité de bien structurer et de bien ordonner les différentes étapes de segmentation et d'identification (figure 4). Cependant en raison de sa simplicité nous conserverons ce schéma comme base pour la suite de l'exposé.

.../...

NUMERO DU SEGMENT	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	COMMENTAIRES
PHRASE IDEALE		a	k	o	m	ə			a	n		a	t	o	l	
PHRASE RECONNUE	p k t	a ɛ̃ ɑ̃		o ø r	b m g n	ɛ̃ a	f r	p k t	a ɛ̃	n b d v	n w d ɲ	ɛ̃ ɛ a	t p k	ø o œ r	v n z m	CHOIX N° 1 CHOIX N° 2 CHOIX N° 3 CHOIX N° 4
N° DE LA SOLUTION CORRESPONDANT AU PHONEME IDEAL	+	1	-	1	2		+	+	1	1	+	3	1	2		+=INSERTION -=OMISSION =NON RECONNU
PROBABILITE AFFECTEE AU SEGMENT	.78	1		1	1	1	1	.50	1	1	1	1	1	1	1	
PLOSIF	+				.50		-	+		.50	.50		+		-	TRAIT DETECTE:+
FRICATIF	-				-		-	-		-	-		-		-	TRAIT NON DETECTE : -
VOISE	-				+		-	-		+	+		-		.92	0 VALEUR 1=PRESENCE DU TRAIT POSSIBLE
N° DE LA CLASSE	1	V		V	-3	V	12	1	V	-3	-3	V	1	V	12	VOYELLE =V; la signification du n° de la consonne est donnee dans la figure

FIG.3 SORTIES DE L'ANALYSEUR PHONETIQUE POUR LA PHRASE: *A COMME ANATOLE*

OUTPUT OF THE PHONETIC ANALYZER TOGETHER WITH THE DETECTION OF THE CLASS OF PHONEMES FOR THE SENTENCE ""*A COMME ANATOLE*"

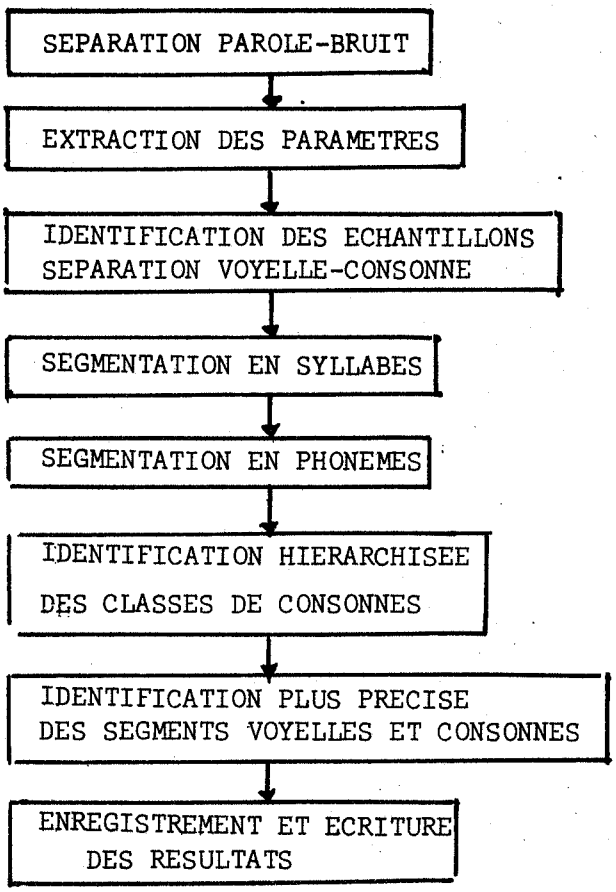


FIG.4 DIFFERENTES ETAPES DE L ANALYSEUR PHONETIQUE

MAIN STEPS OF THE PHONETIC ANALYZER

N° DE LA CLASSE	SIGNIFICATION	CONTENU
0	CONSONNES	
1	PLOSIVES SOURDES	p t k
2	PLOSIVES VOISEES	b d g
3	*FRICATIVES*	f s ʃ ʒ j r ɲ
-3	NON f s ʃ ʒ	
4		ʒ r ɲ j
-4	FRICATIVE SOURDE+ r	f s ʃ r
8	VOISE	
-8	SOURD	f s ʃ p t k
11	PLOSIVE	p t k b d g
12	NON PLOSIF	

FIG.5 PRINCIPALES CLASSES DE PHONEMES DETECTEES PAR *SONIADEL*

MAIN CLASSES USED BY "SONIADEL"

III - EXTRACTION DES PARAMETRES DE BASE

Ces paramètres faciles à déterminer sont ensuite utilisés comme indices pour la segmentation ou pour l'identification de certains traits.

Les paramètres de départ sont évidemment les sorties des 14 filtres et le fondamental F_0. A partir de ces paramètres on calcule (cf annexe 1):

- L'énergie du signal toutes les 13,3 ms (indice de segmentation en syllabe) : E (t)

- La dérivée du signal (pente entre deux échantillons successifs) qui est un indice de transition entre phonèmes car il mesure l'évolution du spectre : P (t)

- Le centre de gravité en fréquence du spectre à l'instant t (pour chaque échantillon). Cet indice permet en particulier de détecter les fricatives non voisées : G (t)

- La variance V (t) du spectre autour de sa valeur moyenne et son relief R (t), paramètres indiquant si le spectre est un spectre chahuté ou plat (permettant ainsi de caractériser certaines fricatives).

- La position des formants (paramètre en cours d'étude : ROSSI, LE CORRE, 1978).

Outre ces paramètres de base, d'autres à usage plus limité comme la valeur de l'énergie dans une zone de fréquence donnée sont calculés progressivement pour mettre en évidence certains phénomènes plus particuliers.

IV - SEGMENTATION

IV.1 - La détection du début et de la fin de parole

Pour cette détection, on cherche les indices permettant de caractériser le signal de parole par rapport au bruit ambiant (sous toutes ses formes).

Or, le signal de parole est caractérisé par :

- Une énergie supérieure à un certain niveau
- par la présence obligatoire du fondamental pendant une certaine durée (si du moins la parole n'est pas chuchotée).
- par un spectre présentant des formants au cours du temps
- par la présence d'énergie dans les basses fréquences (caractéristique des voyelles et des consonnes voisées).

.../...

Si certains bruits présentent des caractéristiques voisins du signal de parole (musique, sifflement, ...) on peut malgré tout considérer que, si l'une de ces caractéristiques est absente pendant une durée déterminée, la probabilité de présence de parole devient faible.

Les paramètres utilisés pour mesurer les 5 caractéristiques précédentes sont respectivement l'amplitue E (t), le relief du spectre R (t) (pour mesurer la présence de formants), l'évolution temporelle P (t), le nombre d'échantillons sonores (dont le fondamental est compris dans la plage de variation normale d'un locuteur humain) l'énergie moyenne de chaque échantillon sur un intervalle de temps donné et l'énergie moyenne dans le premier canal sur le même intervalle de temps.

Ces paramètres sont utilisés dans un premier filtre au moment de l'enregistrement pour éliminer les signaux de bruit et arrêter l'enregistrement de parole lorsque le silence dépasse une certaine durée.

Faute de temps, une évaluation objective de ses performances n'a pas été effectuée. On peut considérer que par expérience, dans l'ensemble il donne satisfaction.

A la suite de ce premier filtre, pendant l'analyse phonétique, le signal enregistré est examiné par un nouveau filtre dont le rôle essentiel est de mieux localiser les débuts et fins d'enregistrements parfois précédés ou suivis de certains bruits (froissements de papier, bruits de respiration, etc...), dont le spectre peut être confondu avec celui des fricatives sourdes.

Par exemple, on cherche à savoir si entre ces bruits et le début (ou la fin) effectifs de la parole voisée on n'a pas un temps de silence trop long. Dans ce cas, le spectre ne correspondrait pas à une fricative ou à une plosive non voisée.

IV.2 - La segmentation en syllabes

- Méthode

Le principe de base de cette segmentation est la recherche du noyau vocalique de la syllabe. Celle-ci s'effectue en deux étapes :

- Le premier paramètre utilisé est l'amplitude E (t) qui nous permet de décomposer le signal de parole en segments consécutifs dont les frontières correspondent aux minima E_{mi} de la courbe E (t) et dont les centres correspondent aux maxima E_{mi} de cette courbe.

Si la différence entre le maximum E_{mi} et le minimum E_{mi} est inférieure à un seul s, le sommet détecté n'est pas jugé significatif.

Au cours de la deuxième étape, ces segments sont réexaminés un à un. Une fonction linéaire de 9 paramètres est calculée. Si cette fonction dépasse un seuil s', une syllabe est détectée sinon le segment considéré est relié au segment suivant pour former un nouveau segment et éventuellement une nouvelle syllabe (GRESSER, MERCIER, 1974).

.../...

Ces 9 paramètres sont les suivants :

- La différence d'énergie entre le maximum du segment et les minimas voisins
- L'énergie moyenne de chaque échantillon
- L'énergie totale du segment
- L'énergie du maximum
- Le nombre d'échantillons voisés
- Le nombre d'échantillons étiquetés "voyelle"
- La longueur du segment
- L'évolution spectrale
- Le nombre d'échantillons sourds ou "consonne" entourant les échantillons "voyelle" du segment.

- Caractérisation de la syllabe

A l'issue de cette segmentation, chaque syllabe est représentée par les paramètres suivants : adresses de début et de fin de syllabe, longueur, intensité de la syllabe, valeur du fondamental F_0 au centre de la voyelle, localisation du noyau vocalique de la syllabe. Tous ces paramètres sont ensuite utilisés par le programme d'analyse prosodique.

- Résultats

Ceux-ci sont résumés par la figure 6

LOCUTEUR	1	1	2	3	4	5	TOTAL	POURCENT
NOMBRE DE PHRASES	23	17	31	9	27	27	134	
NOMBRE DE SYLLABES	106	207	143	108	123	145	832	
OMISSION	1	3	2	2	0	1	9	1.08%
INSERTION	5	3	10	0	2	7	27	3.24%

FIG.6 RESULTATS DE LA SEGMENTATION EN SYLLABES
RESULTS OF THE SEGMENTATION INTO SYLLABLES

.../...

On distingue deux types d'erreurs : l'omission ou l'insertion de syllabes.

Les principales erreurs d'omission sont dues à la non détection (dans certaines conditions) de la voyelle /i/ ou de la voyelle /u/ (pas de maximum sur la fonction énergie E (t)) ou encore à la non détection de certaines syllabes non accentuées (de_la, ma_dame...) prononcées très rapidement.

En cas d'insertion d'une syllabe supplémentaire, 2 cas peuvent se produire :

- tantôt les consonnes voisées telles que /r/, /l/ ou /z/ ayant une énergie assez élevée peuvent être prises pour des voyelles en particulier dans les contextes gr bl br gl ou ji ju...

- tantôt la voyelle de la syllabe en particulier dans le contexte [voyelle] suivie de /r/ est coupée en deux segments (détection de 2 maximas d'énergie).

Remarque :

Dans la figure 6, on n'a pas considéré comme erreur le fait d'insérer ou d'omettre un e muet après certaines consonnes en fin de mot comme par exemple dans les mots "comme" ou "Anatole" de la phrase "A comme Anatole".

IV.3 - La segmentation en phonèmes

- Localisation des voyelles

La voyelle est en principe localisée autour du maximum d'énergie de la syllabe. En pratique, on calcule le début et la fin de la zone stable de la syllabe, le début et la fin des échantillons étiquetés : "voyelle". Ce qui est considéré comme zone voyelle et un compromis entre ces deux zones. Les principaux indices utilisés pour les définir sont la pente du signal P (t) et la localisation des formants dans le spectre à chaque instant t, paramètre permettant d'affecter une étiquette de voyelle ou de consonne à chaque échantillon (cf paragraphe V.I).

On a parfois des erreurs de localisation comme par exemple dans la syllabe ji où le maximum d'énergie se trouve sur des échantillons du phonème / ʒ/ qui ont été étiquetés "voyelle".

- Localisation des consonnes

Cette localisation est plus difficile et plus complexe que la localisation des voyelles. Plusieurs étapes sont nécessaires ; la plus importante est la première que nous présentons ici.

Entre chaque zone voyelle, on considère qu'il peut y avoir une, plusieurs ou aucune consonne. Pour les repérer on utilise la pente P(t) qui permet de détecter les zones stables, chacune d'elle pouvant être le centre d'une consonne stationnaire (fricative, nasale ou liquide) ou le centre du silence précédant l'explosion d'une occlusive.

.../...

A l'aide d'un seuil s sur le fonction P (t) on détecte en plus, le début et la fin des zones instables, les pentes moyennes et maximales de ces zones instables, l'énergie minimale de chaque zone stable etc... c'est-à-dire un ensemble de paramètres utilisés par la suite comme indices de détections de certains phonèmes (plosives par exemple).

A ce stade, une première procédure d'identification de classe de phonèmes est appliquée à chaque zone stationnaire ainsi repérée.

Les résultats de la segmentation sont résumés dans la figure 7.

	N %	N %	N %	N %	N %
LOCUTEUR	1	1	2	3	4
TOTAL	340	356	346	238	276
OMISSION	16 4.7	24 6.74	36 10.4	15 6.3	21 7.60
INSERTION	32 9.4	21 5.89	44 12.7	3 1.26	23 8.33

FIG.7 RESULTATS DE LA SEGMENTATION EN PHONEMES

N=NOMBRE DE PHONEMES

RESULTS OF THE SEGMENTATION INTO PHONEMES

Remarque :

Les consonnes /r/ et dans une moindre mesure /p/ dans le contexte [voyelle] suivi de /r/ ou /l/ ainsi que les semi-voyelles /w/, /j/ sont parfois incluses dans la voyelle voisine, entraînant ainsi de fréquentes erreurs d'omission. Les autres sources d'omission apparaissent fréquemment dans les contextes /C/ /C/ ou diphtongue /V/ /V/ ou /SV/ /V/ où /SV/ est une semi-voyelle, /V/ une voyelle et /C/ une consonnne.

V - IDENTIFICATION DES PHONEMES

Une procédure d'identification hiérarchisée est utilisée :

V.1 - Distinction voyelle - consonne

Méthode

La première identification consiste à classer chaque échantillon en voyelle, consonne ou silence.

.../...

Pour cette identification, on calcule les paramètres suivants :

- Localisation et mesure du maximum d'énergie dans les basses fréquences (entre 250 et 650 Hz)

- Localisation et mesure du maximum d'énergie dans les hautes fréquences (entre 2800 et 4200 Hz)

- Localisation et mesure du maximum d'énergie dans les fréquences centrales.

- Localisation et mesure du premier antiformant (minimum d'énergie dans le spectre)

On applique quelques tests simples sur ces paramètres et en fonction des résultats de ces divers tests, l'échantillon est étiqueté : voyelle ou consonne.

Résultats : sur les 238 segments phonétiques extraits des 9 phrases du locuteur 3 , seuls 9 segments consonantiques ont été étiquetés "voyelle". Ces consonnes étaient les suivantes: /l/, /m/, /v/, /r/, /w/

Remarque : Cette procédure demande sûrement à être affinée pour éviter ces erreurs par une utilisation plus systématique des résultats acquis sur l'étude des formants.

V.2 - Identification des voyelles

Méthode

Chaque échantillon des segments "voyelle" est identifié globalement à l'aide de fonctions linéaires des énergies dans chaque zone de fréquence (filtres du vocodeur). Un apprentissage est nécessaire pour déterminer les coefficients de ces fonctions linéaires de séparation. Enfin une valeur moyenne est calculée sur l'ensemble des échantillons de la zone voyelle.

Résultats

Ceux-ci sont résumés dans la matrice de confusion entre voyelles de la figure 8. Cette figure nous donne les résultats de reconnaissance pour un ensemble de 11 voyelles, celles-ci pouvant également être confondues avec le phonème /r/ classé parmi les voyelles et les consonnes.

La principale source d'erreur provient de la présence d'un grand nombre de /ə/ muets dans l'ensemble des phrases prononcées par les 4 locuteurs. D'autre part, la position de ce phonème au centre du "triangle des voyelles" et les diverses formes qu'il peut présenter favorisent sa confusion avec les autres phonèmes.

.../...

La deuxième principale source d'erreur est la confusion / a/, / $\tilde{\varepsilon}$/

En règle générale on peut considérer que les confusions des phonèmes se font avec les phonèmes voisins dans le triangle des voyelles situé dans le plan F_1, F_2 (DELATTRE, 1965).

Pour les 574 voyelles présentées dans la figure 8, le pourcentage de reconnaissance (1er choix) des voyelles est de 47,21%. Si on fait abstraction des résultats sur les phonèmes / ə/, / œ /, /ø /. Ce pourcentage passe à 55,06% (pour 4 locuteurs masculins dont le locuteur d'apprentissage).

phoneme prononce

reconnu	i	e	ɛ	a ɑ	ɔ o	u	y	ø,œə	ɛ̃	ɑ̃	õ	r
i	26	3					3	1			2	
e	12	59	14	1			4	1				
ɛ		5	32	7				2	4			
a ɑ		2	24	35	9			4				
ɔ o	2	3		3	43	6		51			3	
u	3	1		1		7		4				
y	7	4					15	8				
ø œ	1	2		5	18		2	10	2			
ɛ̃			7	40				14	14			
ɑ̃				8				1		7		
õ				2		6		1		9	9	
r								3		2		

FIG.8 MATRICE DE CONFUSION ENTRE VOYELLES

CONFUSION MATRIX FOR THE VOWELS

.../...

V.3 - Identification des consonnes

Pour la reconnaissance des consonnes, avant d'appliquer la méthode globale précédente, on utilise donc une classification hiérarchisée. Cette procédure permet en particulier d'améliorer la classification de quelques consonnes comme les plosives caractérisées par des transitions rapides et d'autre part de prendre en compte le contexte c'est-à-dire les résultats déjà acquis sur les phonèmes voisins.

Au sommet de cette hiérarchie on trouve les classes les plus faciles à détecter : plosives, fricatives, voisées, sourdes.

V.3.1 - Séparation sourde - sonore

On utilise les indices suivants :

- les sorties du détecteur de fréquence fondamentale
- L'énergie dans la bande de fréquence 250 - 450 Hz
- Les différences d'énergie entre le 1er canal et les deux canaux suivants
- Un test sur l'étiquette des échantillons
- La différence des valeurs du fondamental entre deux échantillons consécutifs.

Les résultats sont résumés dans la figure 9. Dès que le trait de voisement (respectivement de non-voisement) n'est pas détecté sur un échantillon du segment de la consonne sonore (respectivement sourde), on ne prend pas de décision. La plupart des erreurs proviennent d'une succession de deux consonnes dont l'une est sonore et l'autre sourde et donc du déplacement du trait de l'une sur l'autre.

REPONSE	N	%
CORRECTE	263	79.45
PAS DE DECISION	56	16.9
ERREUR	12	3.62
TOTAL	331	100

FIG.9A DETECTION DU TRAIT DE VOISEMENT POUR LES CONSONNES SONORES

DETECTION OF VOICED CONSONANTS

REPONSE	N	%
CORRECTE	153	78.86
PAS DE DECISION	39	20.1
FAUSSE	2	1.03
TOTAL	194	100

FIG.9B DETECTION DU TRAIT DE NON-VOISEMENT POUR LES CONSONNES SOURDES

DETECTION OF UNVOICED CONSONANTS

FIG.9

.../...

V.3.2. Détection des plosives

Le critère principal est l'instabilité très forte du spectre avant la voyelle ou la consonne suivante ainsi que la présence dans certains cas d'un maximum d'énergie au voisinage immédiat du début de l'explosion. Enfin cette explosion est très souvent précédée d'une zone de faible intensité.

Les résultats sont résumés dans la figure 10.

La plupart des erreurs apparaissent dans le cas de consonnes nasales parfois prises pour des plosives ; on peut y ajouter aussi la fricative voisée /v/.

Certaines plosives voisées donnent naissance à des erreurs, soit parce-que l'explosion n'est pas assez forte, soit parce-qu'elle n'est pas précédée d'une zone de silence (minimum d'énergie trop élevé).

Les critères utilisés pour séparer les plosives voisées des autres consonnes voisées sont les suivants :

- Energie du minimum faible
- présence d'un silence avant l'explosion
- énergie moyenne au-dessus de 450 Hz faible
- Temps très court entre l'échantillon correspondant au minimum d'énergie et le début de l'explosion.
- Energie de voisement avant l'explosion peu élevée
- Le minimum d'énergie dans le premier canal se trouve après l'instant où l'énergie totale est minimale.

REPONSE	N	%
CORRECTE	160	80.
PAS DE DECISION	29	14.5
ERREUR	11	5.5
TOTAL	200	100.0

FIG.10a:RESULTATS SUR LES CONSONNES PLOSIVES

REPONSE	N	%
CORRECTE	99	84.6
PAS DE DECISION	11	9.4
ERREUR	7	6.0
TOTAL	117	100.0

FIG.10b:RESULTATS SUR LES CONSONNES NON PLOSIVES

FIG.10: SEPARATION DES CONSONNES PLOSIVES ET NON PLOSIVES
PLOSIVE-NON PLOSIVE DISCRIMINATION

.../...

V.3.3. Séparation plosive-absence de consonne en début de mot

Les 5 indices ci-dessous sont utilisés pour distinguer la présence d'une plosive sourde de l'absence des plosives.:

- L'existence d'une zone de transition très rapide
- La présence pendant l'explosion d'un minimum de la courbe E (t) (cas fréquent pour la plosive /k/ ou pour les suites /t/ / r/ ou / k/ /r /
- La présence d'une énergie élevée dès le début de la transition
- Une différence sensible entre le centre de gravité des échantillons du début (transition) et des échantillons de la zone stable de la voyelle qui suit.
- Une énergie moyenne (pondérée) des 3 derniers canaux, élevée, au moment de l'explosion.

Résultats

Sur 72 mots commençant par une plosive, le poids moyen affecté à ce segment à l'aide de ces indices est de 0,72 (le poids minimal étant de 0 pour l'absence de consonne et de 1 pour la présence d'une consonne).

Sur 39 mots commençant par une voyelle on a rajouté un segment supplémentaire dont le poids moyen est de 0,30.

Il est évident que ces résultats ne sont pas encore satisfaisants et qu'il faudra modifier le calcul du poids et trouver de nouveaux indices.

V.3.4. Détection des "fricatives" (sourdes)

Le critère essentiel est le suivant : si le nombre d'échantillons dont le centre de gravité est au-dessus d'un seuil α est suffisant et si le voisement est nul on détecte la présence d'une fricative sourde /f /, /s / / ʃ / ou éventuellement / r/ après une plosive telle que /t /. Dans cette classe qui correspond à la classe 3 de la figure 5 on peut en outre détecter les phonèmes suivants / ʒ/, / j/ ou /ɲ /, mais dans ce cas le critère de voisement permet de les distinguer des fricatives sourdes.

Les résultats sont résumés dans la figure 11.

Les 4 erreurs sont dues à deux phonèmes / z/, un phonème / k/ et un / t/ suivi d'un e muet.

Cet indice semble donner des résultats satisfaisants malgré sa simplicité.

.../...

REPONSE	N	%
CORRECTE	54	74.0
PAS DE DECISION	19	26.0
ERREUR	0	0
TOTAL	73	100.

FIG.11a:DETECTION DE f s ʃ

REPONSE	N	%
CORRECTE	109	76.22
PAS DE DECISION	30	20.97
ERREUR	4	2.79
TOTAL	143	100

FIG.11b:DETECTION DE LA CLASSE:-3

FIG.11 DETECTION DES FRICATIVES SOURDES

DETECTION OF UNVOICED FRICATIVES

V.3.5. Détection de / ʀ/ ou / l/ après une plosive

Les indices suivants sont utilisés :

Après la plosive, on recherche un maximum d'énergie suivi d'assez près d'un minimum. On remarque également la présence d'un centre de gravité assez élevé et d'une énergie relativement forte dans la zone 450 - 1300 Hz.

Sur 12 segments qui sans l'utilisation de ces indices auraient été classés / f/, / s/, / ʃ/, 6 ont été classés /ʀ / ou / l/ et ainsi récupérés.

V.3.6. Identification définitive

Pour distinguer les phonèmes à l'intérieur de chaque classe on fait appel à une méthode identique à celle utilisée pour le classement des voyelles (Paragraphe V.2). Pour chaque échantillon un ensemble de fonctions linéaires des énergies dans chaque canal est calculé. On détermine ensuite la valeur moyenne sur l'ensemble des échantillons représentant la consonne.

Les résultats pour l'ensemble des phonèmes (voyelles + consonnes) sont résumés dans le tableau de la figure 12, pour 4 locuteurs masculins, le premier, correspondant au locuteur d'apprentissage. Comme on le remarque sur les figures 2 et 3 plusieurs réponses sont possibles ; aussi avons-nous présenté les résultats en tenant compte de cette possibilité. La ligne correspondant au choix N° 1 indique que la première solution fournie par l'analyseur "soniadel" correspond au "bon choix", la ligne correspondant au choix N° 2 indique que le phonème correct a été identifié dès la 2ème réponse et on s'aperçoit ainsi que sur 31 phonèmes possibles, pour le locuteur d'apprentissage 1, environt 85% entre eux sont reconnus dès les 3 premièressolutions, pourcentage qui descend à 75% pour l'ensemble des 3 autres locuteurs masculins. Il semblerait également que les résultats se dégradent en fonction de la longueur de la phrase (mots isolés → phrases courtes → phrases longues) mais ceci demanderait à être vérifié de façon plus systématique.

..../...

	N	%	N	%	N	%	N	%	N	%
LOCUTEUR	1		1		2		3		4	
CHOIX N° m=1	118	52.67	131	39.33	132	42.58	75	33.63	94	36.86
CHOIX N° m=1,2	172	76.78	237	71.38	197	63.54	130	58.29	155	60.78
CHOIX N° m=1,2,3	198	88.39	275	82.83	232	74.83	164	73.54	200	78.43
IDENTIFIE	209	93.30	293	88.25	258	83.22	194	87.0	219	85.9
NON IDENTIFIE	15	6.70	39	11.74	52	16.78	29	13.0	36	14.1
TOTAL	224	100.	332	100.	310	100.	223	100.	255	100.
NOMBRE DE PHRASES	23		17		31		9		27	

FIG.12: POURCENTAGE DE RECONNAISSANCE DES PHONEMES
PERCENTAGE OF PHONEME RECOGNITION
N= NOMBRE DE PHONEMES

CONCLUSION :

Nous avons décrit les principales fonctions de l'analyseur phonétique "soniadel", les indices utilisés, et présenté les résultats de segmentation et de classification. Si les pourcentages de détection sont encourageants (85%), le pourcentage de reconnaissance ne semble pas encore suffisant et des recherches dans des directions diverses seront nécessaires : adaptation au locuteur, amélioration du corpus d'apprentissage, choix d'autres fonctions de reconnaissance, recherche de nouveaux indices, utilisation d'autres analyses acoustiques et de règles phonologiques.

REMERCIEMENTS :

Ce travail a été effectué au sein du groupe communication homme-machine du C.N.E.T. et de ce fait a bénéficié de la collaboration et de l'aide permanente des membres de ce groupe (en particulier P. QUINTON, R. VIVES, I. EL MALLAWANY, C. LE CORRE). De nombreuses autres personnes ont également apporté leur contribution à ce travail tant du point de vue logiciel, matériel qu'administratif ; je remercie en particulier Mme ARZUL pour le travail de dactylographie.

BIBLIOGRAPHIE

P. DELATTRE : Comparing the phonetic features of english german spanish and french. Julius Groos Verlag 1965

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 - Academic Press 1975 p 359 à 382

G. MERCIER, P. QUINTON, R. VIVES : Dialogue homme-machine avec Keal Recherches acoustiques volume IV 1978

M.T. ROTH, J. VAISSIERE : Etude des variations acoustiques de la voyelle dans les monosyllabes en français 9èmes J.E.P. LANNION 1978

M. ROSSI, C. LECORRE : Indices de détection de formants sur l'analyse spectrale par canaux - 9èmes J.E.P. LANNION 1978

J. VAISSIERE : Automatic segmentation of continuous speech in words in french 9th I.C.A. MADRID 4 - 9 juillet 1977 p 485

R. VIVES L'analyse lexicale dans le système KEAL pour la reconnaissance de la parole continue - 7èmes J.E.P. NANCY 19 - 21 mai 1976 p 115 à 127

A N N E X E

Définition des principaux paramètres

1. - L'énergie dans la bande de fréquence correspondant au canal j à l'instant t : canal (j, t).

2. - Le nombre de bandes de fréquence (canaux du vocodeur) = n canaux

3. - L'énergie du spectre à l'instant t : $E(t) = \sum_{j=1}^{n\ canaux} canal\ (j,t)$

4. - La valeur moyenne de l'énergie dans chacun des canaux à l'instant t :

$$moy\ (t) = \frac{E(t)}{n\ canaux}$$

5. - La pente à l'instant t entre deux échantillons consécutifs : P (t)

$$P(t) = \sum_{j=1}^{n\ canaux} \left| canal\ (j, t) - canal\ (j, t+1) \right|$$

6. - La variance de l'énergie de chaque canal autour de la valeur moyenne moy (t) : Var (t) =

$$\sum_{j=1}^{n\ canaux} (canal\ (j, t) - moy\ (t))^2$$

7. - Le centre de gravité en fréquence de l'échantillon t : G (t)

$$G(t) = \frac{\sum_{j=1}^{n\ canaux} j . canal\ (j, t)}{\sum_{j=1}^{n\ canaux} canal\ (j, t)}$$

.../...

8. - Le relief du spectre (répétition spectrale) : R (t)

$$R(t) = \sum_{j=1}^{n \text{ canaux } -1} \left| \text{canal } (j+1, t) - \text{canal } (j, t) \right|$$